SLPAT 2015

6th Workshop
on
Speech and Language Processing
for Assistive Technologies
(SLPAT)

Workshop Proceedings

11 September, 2015
Dresden, Germany

Order print-on-demand copies from:

Curran Associates
57 Morehouse Lane
Red Hook, New York 12571
USA
Tel: +1-845-758-0400
Fax: +1-845-758-2633
curran@proceedings.com

ISBN 978-1-941643-79-2
ISSN 2411-9962

Introduction

We are pleased to bring you the Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), held in Dresden, Germany, on 11 September, 2015. We received 35 paper submissions, of which 22 were chosen for oral presentation and one was a system demonstration; all 23 papers are included in this volume.

This workshop was intended to bring researchers from all areas of speech and language technology with a common interest in making everyday life more accessible for people with physical, cognitive, sensory, emotional or developmental disabilities. This workshop builds on five previous such workshops (co-located with conferences such as ACL, NAACL, EMNLP and Interspeech); it provides an opportunity for individuals from research communities, and the individuals with whom they are working, to share research findings, and to discuss present and future challenges and the potential for collaboration and progress.

While Augmentative and Alternative Communication (AAC) is a particularly apt application area for speech and natural language processing technologies, we purposefully made the scope of the workshop broad enough to include assistive technologies (AT) as a whole, even those falling outside of AAC. Thus we have aimed at broad inclusivity, which is also manifest in the diversity of our Program Committee. We are also very delighted to have Prof. Jonas Beskow from the Royal Institute of Technology, Stockholm, Sweden, as invited speaker.

The success of this SLPAT 2015 edition was due to the authors who submitted such interesting and diverse work and which generated so intense discussions. Finally, we must thank all the people who made this event possible: The members of the Program Committee for completing their reviews promptly, and for providing useful feedback for deciding on the program and preparing the final versions of the papers. The Interspeech organisers who, in many ways, made the organisation easier, the ISCA Administrative Secretariat for handling finance and the Dresden University of Technology which hosted the event in their premises.

Jan Alexandersson, Ercan Altinsoy, Heidi Christensen,
Peter Ljunglöf, François Portet, and Frank Rudzicz

Co-organizers of SLPAT 2015

List of People

Organizers:

Jan Alexandersson, German Research Center for Artificial Intelligence, Germany.
Ercan Altinsoy, Technical University of Dresden, Germany.
Heidi Christensen, University of Sheffield, UK.
Peter Ljunglöf, University of Gothenburg, Sweden.
François Portet, University of Grenoble Alpes, France.
Frank Rudzicz, University of Toronto, Canada.

Program committee:

Jean-Yves Antoine, Université François Rabelais de Tours, France.
John Arnott, University of Dundee, UK.
Melanie Baljko, York University, Canada.
Stefan Bott, Universität Stuttgart, Germany.
Annelies Braffort, LIMSI-CNRS, France.
Corneliu Burileanu, University Politehnica of Bucharest, Romania.
Heriberto Cuayahuitl, Heriot-Watt University, UK.
Stuart Cunningham, University of Sheffield, UK.
Rickard Domeij, Swedish Language Council, Sweden.
Michael Elhadad, Ben Gurion University, Israel.
Isabelle Estève, University of Grenoble Alpes, France.
Corinne Fredouille, CERI/LIA – University of Avignon, France.
Kallirroi Georgila, University of Southern California, USA.
Stefan Goetze, Fraunhofer Institute for Digital Media Technology, Germany.
Björn Granström, KTH, Sweden.
Phil Green, University of Sheffield, UK.
Mark Hasegawa-Johnson, University of Illinois at Urbana-Champaign, USA.
Per-Olof Hedvall, Lund University, Sweden.
Matt Huenerfauth, City University of New York, USA.
Per Ola Kristensson, University of Cambridge, UK.
Benjamin Lecouteux, University of Grenoble Alpes, France.
Greg Lesher, DynaVox Technology, USA.
William Li, Massachusetts Institute of Technology, USA.
Eduardo Lleida, University of Zaragoza, Spain.
Ornella Mich, Fondazione Bruno Kessler, Italie.
Climent Nadeu, Universitat Politècnica de Catalunya, Spain.
Torbjørn Nordgård, Lingit AS, Norway.
Ehud Reiter, University of Aberdeen, UK.
Brian Roark, Google, USA.
Bitte Rydeman, Lund University, Sweden.
Rubén San-Segundo Hernández, Universidad Politécnica de Madrid, Spain.
Michel Vacher, Laboratoire LIG, équipe GETALP, France.
Keith Vertanen, Michigan Tech, USA.
Nadine Vigouroux, IRIT, France.
Ravichander Vipperla, Nuance Communications, UK.
Maria Wolters, University of Edinburgh, UK.

Table of Contents

Invited talk

Paper session 1: Sign language

Paper session 2: Signal enhancement and speech recognition

Paper session 3: Communication aids, speech synthesis and other topics

System demonstration

Talking Heads, Signing Avatars and Social Robots
Exploring multimodality in assistive applications

Jonas Beskow

KTH, Stockholm, Sweden
beskow@kth.se

Abstract

Over the span of human existence our brains have evolved into sophisticated multimodal social signal processing machines. We are experts at detecting and decoding information from a variety of sources and interpreting this information in a social context. The human face is one of the most important social channels that plays a key role in the human communication chain. Today, with computer animated characters becoming ubiquitous in games and media, and social robots starting to bring human-like social interaction capabilities into the physical world, it is possible to build applications that leverage the unique human capability for social communication new ways to assist our lives and support us in a variety of domains.

This talk will cover a series of experiments attempting to quantise the effect of several traits of computer generated characters/robots such as visual speech movements, non-verbal signals, physical embodiment and manual signing. It is shown that a number of human functions ranging from low-level speech perception to learning can benefit from the presence of such characters when compared to unimodal (e.g. audio-only) settings. Two examples are given of applications where these effects are exploited in order to provide support for people with special needs – a virtual lipreading support application for hard of hearing and a signing avatar game for children with communicative disorders.

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, page 1,
Dresden, Germany, 11 September, 2015.

Bridging the gap between sign language machine translation and sign language animation using sequence classification

Sarah Ebling[1], *Matt Huenerfauth*[2]

[1]Institute of Computational Linguistics
University of Zurich
Zurich, Switzerland
ebling@cl.uzh.ch
[2]Rochester Institute of Technology (RIT)
Golisano College of Computing and Information Sciences
Rochester, NY, USA
matt.huenerfauth@rit.edu

Abstract

To date, the non-manual components of signed utterances have rarely been considered in automatic sign language translation. However, these components are capable of carrying important linguistic information. This paper presents work that bridges the gap between the output of a sign language translation system and the input of a sign language animation system by incorporating non-manual information into the final output of the translation system. More precisely, the generation of non-manual information is scheduled after the machine translation step and treated as a sequence classification task. While sequence classification has been used to solve automatic spoken language processing tasks, we believe this to be the first work to apply it to the generation of non-manual information in sign languages. All of our experimental approaches outperformed lower baseline approaches, consisting of unigram or bigram models of non-manual features.

1. Introduction

Sign languages are often the preferred means of communication of deaf and hard-of-hearing persons, making it vital to provide access to information in these languages. Technologies for automatically translating written text (in a spoken language[1]) into a sign language would therefore increase the accessibility of information sources for many people.

Sign languages are natural languages and, as such, fully developed linguistic systems. While there are a variety of sign languages used internationally, they share several key properties: Utterances in sign languages are produced with the hands/arms (the *manual activity*) and the shoulders, head, and face (the *non-manual activity*). Manual and non-manual components together form the *sublexical components*.

1.1. Sign language production pipeline

While the input to a translation system such as the one outlined above would be a written text, the output is less obvious: Ultimately, the goal would be to produce an animation of a virtual human character performing sign language, i.e., a *sign language avatar*. Most sign language machine translation systems produce some form of symbolic output. In the ideal case, this output should be suitable to serve as the *input* for an animation-synthesis system.

Unfortunately, to date, this sign language production pipeline is often left incomplete, in that the output of many machine translation systems consists of strings of sign language glosses,[2] i.e., information about the manual activity of a signed utterance, only.

This paper presents work that *bridges the gap* between the output of a sign language translation system and the input of a sign language animation system by incorporating non-manual information into the final output of the translation system. More precisely, the generation of non-manual information is scheduled after the machine translation step and treated as a sequence classification task. To our knowledge, this is the first work to apply sequence classification to the generation of non-manual information in sign languages. We show that all of our experimental approaches outperformed lower baseline approaches, including unigram and bigram models of non-manual component sequences.

1.2. Linguistic background and prior work

Experimental research with sign language users has shown that the absence of non-manual information in synthesized signing (sign language animation) leads to lower comprehension scores and lower subjective ratings of the animations [1]. This is because non-manual components in sign languages are capable of assuming functions at all linguistic levels [2]. As an example, in Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS), raised eyebrows are used to express supposition, contrast, or emphasis [3]. A combination of a head movement forward and raised eyebrows is used to mark topicalized constituents. Conditional *if/when* utterances have the head tilt and move forward slightly and the eyebrows go up at the start of the condition part. For rhetorical questions in DSGS, the head tilts and moves forward slightly and the eyebrows are furrowed on the question sign [4].

Non-manual components have been omitted, for example, in a statistical machine translation system that translates between German and German Sign Language [5] and one that

[1] The term *spoken language* refers to a language that is not signed, whether it is represented in its spoken or written form.

[2] Glosses are semantic representations of signs that typically take on the base form of a word in the surrounding spoken language.

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 2–9,
Dresden, Germany, 11 September, 2015.

translates between English, German, Irish Sign Language, and German Sign Language [6]. Massó and Badia [7] took into account mouth morphemes in a statistical machine translation system translating from Catalan into Catalan Sign Language; such mouth movements convey adverbial or aspectual modifications to the meaning of manual signs in that language.

In contrast, in this paper, we deal with multiple types of non-manual components, taking into account the multilinear nature of sign languages. Our work is inspired by linguistic models that represent both the manual and non-manual components of signed utterances [8, 9].

The remainder of the paper is structured as follows: Section 2 introduces the project as part of which the machine translation system is being developed. In particular, the data that served as a basis for the sequence classification experiments is described. In Section 3, we specify our sequence classification approaches, provide further information on the data used for the experiments, explain the experiment configurations, and present as well as discuss the results.

2. Non-manual components in a corpus of DSGS train announcements

We are developing a system that automatically translates written German train announcements of the Swiss Federal Railways into DSGS. Our team includes Deaf[3] and hearing researchers. Example 1 below shows an announcement of the Swiss Federal Railways.

(1) *Ausfallmeldung zur S1 nach Luzern: Die S1 nach Luzern, Abfahrt um 6 Uhr 10, fällt aus.* ('Notice of cancellation regarding the S1 to Lucerne: The S1 to Lucerne, scheduled to leave at 6:10am, has been cancelled.')

The resulting DSGS announcements are presented by means of an avatar. A state-of-the art avatar system, JASigning [11], is used for this. The JASigning character *Anna* is shown in Figure 1.

The train announcements of the Swiss Federal Railways are parametrized in that they are based on templates with slots, where slots are, e.g., the names of train stations, types of trains, or reasons for delays. When automatically translating these announcements, one possibility is to take account precisely of their parametrized nature. However, our goal is to build a translation system that can later be extended to other domains with more lexical and syntactic variation. Hence, a more transferrable translation approach is applied, namely statistical machine translation.

Statistical machine translation systems require parallel corpora as their training, development, and test data. To build a parallel corpus, the Deaf and hearing members of our team manually translated 3000 written German train announcements into DSGS. The DSGS side of the resulting parallel corpus consists of information arranged on three tiers:

1. sign language glosses
2. head, with 13 possible values
3. eyebrows, with 3 possible values

The non-manual components in the DSGS side of our parallel corpus serve various linguistic functions. For example, in our domain of train announcements, we have observed that furrowed eyebrows often occurred during signs with negative polarity, such as the sign BESCHRÄNKEN ('LIMIT'). Raised eyebrows often occurred during signs that express a warning or emphasis, e.g., the signs VORSICHT ('CAUTION') or SOFORT ('IMMEDIATELY'). The syntactic functions mentioned in Section 1.2, topicalization and rhetorical question, also occur frequently in the corpus; a few instances of conditional expressions are also present. Many of these syntactic non-manuals relate to specific words in the sentence (e.g., rhetorical question non-manual components co-occur with question words, such as "WHAT"). Within this paper, we focus on such *lexically-cued non-manuals*. (As discussed in Section 4, we are aware that not all non-manual components are predictable based on the sequence of lexical items in the sentence alone, and we propose to investigate such non-manuals in future work.)

Table 1 shows the DSGS translation of the first part of the train announcement introduced in Example 1, *Ausfallmeldung zur S1 nach Luzern* ('Notice of cancellation regarding the S1 to Lucerne'). Note that the starting and ending times of the non-manual components align with the boundaries of manual activities (as represented through glosses). This has been shown to be the case for non-manual components with linguistic functions; non-manual components that serve purely affective purposes, e.g., expressing anger or disgust, are known to start slightly earlier than the surrounding manual components [12, 13, 14, 15, 16, 17, 18, 19, 20, 21].

3. Generating non-manual information through sequence classification

The goal of our work was to include non-manual information in the process of translating written German train announcements into DSGS. Traditionally, glosses have been the sole representation of sign language in an automatic translation task (cf. Section 1). One way of considering non-manual components in this task is to simply append them to the glosses. This representation is shown in Example 2 for the announcement introduced in Example 1. The non-manual features are printed in bold.

(2) *Ausfallmeldung zur S1 nach Luzern*:
MELDUNG**__Head_forward__Eyebrows_raised**
IX**__Head_back__Eyebrows_raised**
BAHN**__Head_up__Eyebrows_raised**
S1**__Head_down__Eyebrows_raised**
NACH**__Head_up__Eyebrows_neutral**
LUZERN**__Head_up__Eyebrows_raised**
AUSFALL**__Head_down__Eyebrows_raised**

However, such a representation aggravates the issue of data sparseness, since the size of the vocabulary is not only equivalent to the number of unique glosses but to the number of unique combinations of glosses and non-manual features. This increases the likelihood that tokens appear in the decoding phase that have not been seen during training (*out-of-vocabulary items*, OOV). Such a representation also does not accommodate the multimodal nature of sign languages: Three tiers (glosses, head, and eyebrow information) are collapsed into one.

We propose an approach that schedules the automatic generation of non-manual information after the machine translation step and views it as a sequence classification task. This is justified by the fact that the non-manual components in our corpus

[3] It is a widely recognized convention to use the upper-cased word *Deaf* for describing members of the linguistic community of sign language users and, in contrast, to use the lower-cased word *deaf* when describing the audiological state of a hearing loss [10].

Glosses	MELDUNG (‘NOTICE’)	IX (‘IX’)	BAHN (‘TRAIN’)	S1 (‘S1’)	NACH (‘TO’)	LUZERN (‘LUCERNE’)	AUSFALL (‘CANCELLATION’)
Eyebrows	raised				neutral	raised	
Head	forward	back	up	down	up		down

Table 1: DSGS translation of *Ausfallmeldung zur S1 nach Luzern* (‘Notice of cancellation regarding the S1 to Lucerne’)

serve linguistic functions, which means their boundaries align with those of manual components (cf. Section 2). Hence, the process of generating non-manual components can be regarded as a task of labeling glosses (as representations of the manual components) with non-manual features.

Figure 1 visualizes the overall pipeline that transforms a written German train announcement into a DSGS animation: The machine translation system receives as input a German announcement like the one introduced in Example 1. With the help of models learned from our parallel corpus, the system translates the German announcement into DSGS glosses. The glosses in turn serve as input for the sequence classification system. The output of the machine translation and the sequence classification system is then combined and converted into motion data for the avatar. The process of generating the motion data is not illustrated further in the figure, as it is outside of the scope of this paper.

3.1. Conditional Random Fields

Sequence classification has been used to solve various natural language processing problems, such as part-of-speech tagging and chunking (shallow parsing). In contrast to standard classifiers, sequence classifiers are capable of taking into account the sequential nature of data. Sequential Conditional Random Fields (CRFs) [22] are a state-of-the-art approach for this. Given one or more sequences of tokens (the *evidence*), CRFs compute the probability of a sequence of labels (the *outcome*). While multiple evidence layers are permitted, CRFs only allow the prediction of one outcome layer.

The Wapiti toolkit [23] provides an efficient implementation of CRFs.[4] Sequence classification with Wapiti follows a train–test–evaluate cycle. Handcrafted *feature templates* are created to specify which tokens of the evidence are considered for the prediction of the outcome labels. In addition, the *emission order* is declared, indicating whether the evidence is conditioned on label unigrams (emission order 1) or bigrams (emission order 2). During the training step, the feature templates are instantiated with the training data. Wapiti offers a model dump function, which allows the user to investigate the quality of the resulting features.

3.2. Data

To perform the sequence classification experiments in Wapiti, the parallel corpus of 3000 German/DSGS train announcements described in Section 2 was randomly divided into ten folds of 300 sentences each to enable ten-fold cross validation. For each validation round, eight folds were used for training, one was used for development, and one for testing. Using the ground truth as opposed to the machine translation output (cf. Section 2) as data was motivated by our interest in investigating the potential of sequence classification in isolation, without possible error propagation from the preceding machine translation step.

3.3. Experiment configurations

The goal of the experiments described here was to predict the most probable sequence of non-manual features for a sequence of glosses output by the machine translation system (cf. Figure 1). As stated in Section 3.1, CRFs allow the prediction of one outcome layer at a time. Hence, the two label layers head and eyebrows in our corpus (cf. Section 2) could either be collapsed into a single label (Configuration G→H+E, Table 2), or a separate classifier could be trained for each feature (Configurations G→H and G→E, Table 3). A downside of Configuration G→H+E is that there is a potential for data sparseness, as the number of possible outcome labels is equivalent to the number of cross-combinations of head and eyebrow labels occurring in the training data. However, even with this approach, the risk of data sparseness is lower than that of appending the non-manual features to the sign language glosses during the machine translation task, as described at the beginning of Section 3.

Evidence **Gloss**	**Label** **Non-manual**
MELDUNG (‘NOTICE’)	forward_raised
IX (‘IX’)	back_raised
BAHN (‘TRAIN’)	up_raised
S1 (‘S1’)	down_raised
NACH (‘TO’)	up_neutral
LUZERN (‘LUCERNE’)	up_raised
AUSFALL (‘CANCELLATION’)	down_raised

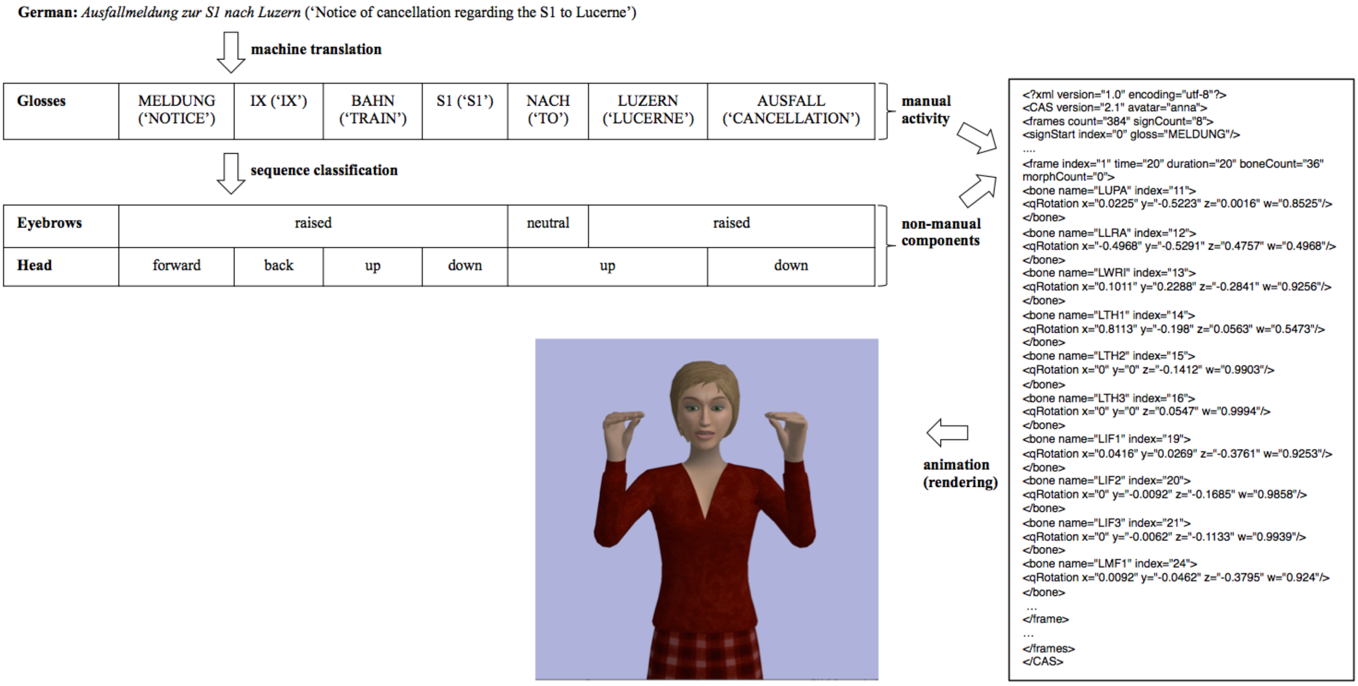
Table 2: Configuration G→H+E

Evidence **Gloss**	**Label** **Head**
MELDUNG (‘NOTICE’)	forward
IX (‘IX’)	back
BAHN (‘TRAIN’)	up
S1 (‘S1’)	down
NACH (‘TO’)	up
LUZERN (‘LUCERNE’)	up
AUSFALL (‘CANCELLATION’)	down

Evidence **Gloss**	**Label** **Eyebrow**
MELDUNG (‘NOTICE’)	raised
IX (‘IX’)	raised
BAHN (‘TRAIN’)	raised
S1 (‘S1’)	raised
NACH (‘TO’)	neutral
LUZERN (‘LUCERNE’)	raised
AUSFALL (‘CANCELLATION’)	raised

Table 3: Configurations G→H (top) and G→E (bottom)

[4] `http://wapiti.limsi.fr/manual.html`



German: *Ausfallmeldung zur S1 nach Luzern* ('Notice of cancellation regarding the S1 to Lucerne')

Glosses	MELDUNG ('NOTICE')	IX ('IX')	BAHN ('TRAIN')	S1 ('S1')	NACH ('TO')	LUZERN ('LUCERNE')	AUSFALL ('CANCELLATION')

Eyebrows	raised				neutral	raised	
Head	forward	back	up	down	up		down

```
<?xml version="1.0" encoding="utf-8"?>
<CAS version="2.1" avatar="anna">
<frames count="384" signCount="8">
<signStart index="0" gloss="MELDUNG"/>
....
<frame index="1" time="20" duration="20" boneCount="36"
morphCount="0">
<bone name="LUPA" index="11">
<qRotation x="0.0225" y="-0.5223" z="0.0016" w="0.8525"/>
</bone>
<bone name="LLRA" index="12">
<qRotation x="-0.4968" y="-0.5291" z="0.4757" w="0.4968"/>
</bone>
<bone name="LWRI" index="13">
<qRotation x="0.1011" y="0.2288" z="-0.2841" w="0.9256"/>
</bone>
<bone name="LTH1" index="14">
<qRotation x="0.8113" y="-0.198" z="0.0563" w="0.5473"/>
</bone>
<bone name="LTH2" index="15">
<qRotation x="0" y="0" z="-0.1412" w="0.9903"/>
</bone>
<bone name="LTH3" index="16">
<qRotation x="0" y="0" z="0.0547" w="0.9994"/>
</bone>
<bone name="LIF1" index="19">
<qRotation x="0.0416" y="0.0269" z="-0.3761" w="0.9253"/>
</bone>
<bone name="LIF2" index="20">
<qRotation x="0" y="-0.0092" z="-0.1685" w="0.9858"/>
</bone>
<bone name="LIF3" index="21">
<qRotation x="0" y="-0.0062" z="-0.1133" w="0.9939"/>
</bone>
<bone name="LMF1" index="24">
<qRotation x="0.0092" y="-0.0462" z="-0.3795" w="0.924"/>
</bone>
 ...
</frame>
...
</frames>
</CAS>
```

Figure 1: Sign language production pipeline: machine translation, sequence classification, and animation

With Configurations G→H and G→E, each label layer (head and eyebrows, respectively) is treated in isolation, which means that dependencies between the two are not captured. However, conceptually, dependencies between these two types of non-manual features exist in that they assume specific linguistic functions together, e.g., topicalization, rhetorical questions, or conditional expressions in DSGS. These dependencies can be accounted for by introducing a cascaded approach, i.e., by using the output of one classifier as additional input for the other. More precisely, the output of the head information classifier can be used as additional evidence for the eyebrow information classifier and vice versa. This is shown as Configurations G_E→H and G_H→E in Table 4. Note that such a representation accommodates the multi-tier nature of sign languages.

To better model the sequential dependencies in a given data set, an IOB representation [24] could be used. In this format, B denotes the first token of a label sequence, I a sequence-internal token, and O is used for tokens that are not part of a sequence of a label under consideration. This format is applied as Configurations G→H_{IOB} and G→E_{IOB} (Table 5). Note that in the case at hand, O does not occur, since the data contains multi-class as opposed to binary annotations and *neutral* is one of the possible class labels.

For our experiments, we applied all of the above seven configurations, as summarized in Table 6.

Among the strengths of CRFs is their ability to handle a large amount of features as well as to cope with redundancy [23]. We provided 26 feature templates for each evidence layer, guided by a template provided by Roth and Clematide [25]. Table 7 lists the 12 context windows used. The table shows that the overall context considered ranged from the three previous tokens to the three following tokens relative to the current position. Each window was included with both emission order 1 (unigram) and 2 (bigram). In addition, raw unigram output distribution and bigram output distribution were included.

Evidence		**Label**
Gloss	**Eyebrows**	**Head**
MELDUNG ('NOTICE')	raised	forward
IX ('IX')	raised	back
BAHN ('TRAIN')	raised	up
S1 ('S1')	raised	down
NACH ('TO')	neutral	up
LUZERN ('LUCERNE')	raised	up
AUSFALL ('CANCELLATION')	raised	down

Evidence		**Label**
Gloss	**Head**	**Eyebrows**
MELDUNG ('NOTICE')	forward	raised
IX ('IX')	back	raised
BAHN ('TRAIN')	up	raised
S1 ('S1')	down	raised
NACH ('TO')	up	neutral
LUZERN ('LUCERNE')	up	raised
AUSFALL ('CANCELLATION')	down	raised

Table 4: Configurations G_E→H (top) and G_H→E (bottom)

3.4. Results

Table 8 shows the results of our experiments obtained using the default settings of Wapiti. "Experimental approach" refers to the configurations described in Section 3.3. The lower baseline for each configuration consisted of using the unigram ("Lower baseline, unigram") and bigram ("Lower baseline, bigram") output distribution of the labels, respectively, i.e., of globally assigning the most frequent label unigram and bigram of the training data, respectively. For each experimental or baseline

Evidence	Label
Gloss	Head
MELDUNG (‘NOTICE’)	B_forward
IX (‘IX’)	B_back
BAHN (‘TRAIN’)	B_up
S1 (‘S1’)	B_down
NACH (‘TO’)	B_up
LUZERN (‘LUCERNE’)	I_up
AUSFALL (‘CANCELLATION’)	B_down

Evidence	Label
Gloss	Eyebrow
MELDUNG (‘NOTICE’)	B_raised
IX (‘IX’)	I_raised
BAHN (‘TRAIN’)	I_raised
S1 (‘S1’)	I_raised
NACH (‘TO’)	B_neutral
LUZERN (‘LUCERNE’)	B_raised
AUSFALL (‘CANCELLATION’)	I_raised

Table 5: Configurations G→H_{IOB} (top) and G→E_{IOB} (bottom)

Configuration	Evidence	Label
G→H+E	glosses	head and eyebrows
G→H	glosses	head
G→E	glosses	eyebrows
G_E→H	– glosses – eyebrows	head
G_H→E	– glosses – head	eyebrows
G→H_{IOB}	glosses	head IOB
G→E_{IOB}	glosses	eyebrows IOB

Table 6: Overview of configurations

Relative position	Description
0	current token
-1 to +1	previous, current, following token
-1 to 0	previous and current token
0 to +1	current and following token
-1	previous token
1	following token
-2 to 0	two previous and current token
-2 to +1	two previous, current, following token
-3 to 0	three previous and current token
-2 to +2	two previous, current, following token
-1 to +2	previous, current, two following tokens
-1 to +3	previous, current, three following tokens

Table 7: Context windows used for the feature templates

approach, Table 8 provides the following numerical information, in analogy to previous work in classification for natural language processing [26, 25, 27]:

- Number of labels
- Token error: This is the mean of the token errors of the ten rounds of a 10-fold cross validation. The token error for an individual validation round is calculated as the percentage of incorrectly predicted labels.
- Standard deviation of token error for the ten rounds
- Confidence interval of token error: This is the confidence interval at a confidence level of 95% calculated over the mean of the token errors using Student’s t-test.
- Sequence error: This is the mean of the sequence errors of a 10-fold cross validation. The sequence error for an individual validation round is calculated as the percentage of incorrectly predicted sequences, i.e., sequences containing at least one token error.
- Standard deviation of sequence error
- Confidence interval of sequence error: This is the confidence interval at a confidence level of 95% calculated over the mean of the sequence errors using Student’s t-test.

The results in Table 8 show that all experimental approaches outperformed their lower baselines (unigram and bigram output distribution); in all cases, the magnitude of the difference was greater than the confidence interval of the values. The error rates of the experimental approaches are notably low, which is at least partly due to the nature of the data used for the experiments: As described in Section 2, the train announcements are highly parametrized in that they are based on a limited set of phrasal templates. As shown in the table, there is a tendency for the bigram baseline to perform better than the unigram baseline, a result that underlines the inherently sequential nature of the data.

3.4.1. Cascaded vs. non-cascaded

Between Configuration G→H (non-cascaded) and G_E→H (cascaded), both predicting head information, Configuration G→H exhibits a lower sequence error rate. Between Configuration G→E (non-cascaded) and G_H→E (cascaded), both predicting eyebrow information, Configuration G_H→E achieved a lower sequence error rate.

To examine the theoretical potential of the cascaded approach, we determined the upper bound, i.e., the result of applying the model learned from the training data on the ground-truth data. In other words, as data for the additional evidence layer (eyebrow information for Configuration G_E→H and head information for Configuration G_H→E), we used the gold-standard annotations of these layers instead of the output of Configurations G→H and G→E, respectively. The resulting numbers are shown in the table as “Upper bound” for Configurations G_E→H and G_H→E: Configuration G_E→H/Upper bound achieved a lower sequence error rate than Configuration G→H. Configuration G_H→E/Upper bound also achieved a lower sequence error rate than Configuration G→E; here, the magnitude of the difference is greater than the confidence intervals of the values. These results show that a cascaded approach is capable of outperforming a non-cascaded approach, and they imply that in DSGS, head information provides more useful information for predicting eyebrow information than vice versa.

Configuration	Labels	Token level			Sequence level		
		Token error (%)	Standard deviation	Confidence interval	Sequence error (%)	Standard deviation	Confidence interval
Predicting H+E	**31**						
Lower baseline, bigram		65.04	0.89	0.63	99.97	0.10	0.07
Lower baseline, unigram		68.16	0.51	0.36	100.00	0.00	n.a.
G→H+E		1.88	0.50	0.36	10.43	2.53	1.81
Predicting H	**13**						
Lower baseline, bigram		63.65	1.05	0.75	99.97	0.10	0.07
Lower baseline, unigram		68.07	0.50	0.36	100.00	0.00	n.a.
G→H		1.62	0.45	0.32	8.96	2.43	1.7
G_E→H		1.62	0.50	0.36	9.19	2.40	1.72
— Upper bound		1.29	0.39	0.28	7.86	2.05	1.46
Predicting E	**3**						
Lower baseline, bigram		20.57	1.07	0.77	94.62	1.80	1.29
Lower baseline, unigram		20.57	1.07	0.77	94.62	1.80	1.29
G→E		0.74	0.24	0.17	6.85	1.81	1.29
G_H→E		0.66	0.16	0.12	5.72	0.96	0.69
— Upper bound		0.45	0.11	0.08	4.21	0.88	0.63
Predicting H$_{\text{IOB}}$	**21**						
Lower baseline, bigram		74.75	3.98	2.85	99.97	0.10	0.07
Lower baseline, unigram		76.41	0.52	0.37	100.00	0.00	n.a.
G→H$_{\text{IOB}}$		1.81	0.56	0.40	9.13	2.84	2.03
Predicting E$_{\text{IOB}}$	**6**						
Lower baseline, bigram		37.07	1.55	1.11	95.79	1.58	1.13
Lower baseline, unigram		43.12	1.48	1.06	98.83	0.48	0.34
G→E$_{\text{IOB}}$		1.41	0.30	0.21	9.96	1.85	1.33

Table 8: Sequence classification experiments: Results

3.4.2. `IOB` vs. non-`IOB`

Between Configuration G→H (non-`IOB` format) and G→H$_{\text{IOB}}$ (`IOB` format), both predicting head information, Configuration G→H produced a lower sequence error rate. Between Configuration G→E (non-`IOB` format) and G→E$_{\text{IOB}}$ (`IOB` format), both predicting eyebrow information, Configuration G→E yielded a lower sequence error rate. In this case, the magnitude of the difference was greater than the confidence interval of the values. These results show that applying an `IOB` format was not beneficial for the task at hand, most likely due to data sparseness: Introducing the `IOB` format doubled the number of labels for Configuration G→E$_{\text{IOB}}$ compared to Configuration G→E (6 vs. 3 labels), while the relative increase was lower for Configuration G→H$_{\text{IOB}}$ compared to Configuration G→H (21 vs. 13 labels), indicating that five head information features appeared sequence-initially only, i.e., spanned over one gloss.

3.4.3. Analysis of features

We examined the 50 highest-weighted (instantiated) features in the models of the experimental approaches of Configurations G→H+E, G_E→H, and G_H→E for the first round of the 10-fold cross validation: Among the highest-weighted features for Configuration G→H+E were 31 bigram features and 19 unigram features. The most frequently occurring feature context window (cf. Table 7) consisted of the current token of the (gloss) evidence layer (i.e., relative position 0). Thus, the identity of a lexical item contributed to the model's prediction of the non-manual feature that co-occurs with it. The second- and third-best performing feature context windows contained the previous token (-1) and the following token (+1) of an evidence layer, respectively, and this was followed by a window containing the current and the following token (0 to +1). Thus, the neighboring lexical items contributed to the prediction of the non-manual feature.

For Configuration G_E→H (predicting head information), the 50 top-weighted features consisted of 26 bigram and 24 unigram features. 48 features used tokens from the gloss evidence layer, while 2 used tokens from the added eyebrow information layer. For Configuration G_H→E (predicting eyebrow information), this number was considerably higher: Among the 50 best-scoring features were 27 that used tokens from the head information layer. Again, this serves as evidence that head information is valuable when predicting eyebrow information in DSGS.

4. Conclusion and outlook

We have presented work that bridges the gap between the output of a sign language machine translation system and the input of a sign language animation system by incorporating non-manual information into the final output of the translation system. This is in contrast to many prior statistical sign language machine translation approaches that did not include non-manual components in their output. The inclusion of such non-manual information enables the final animation-synthesis component of a translation system to control the head and face of a signing avatar.

Our approach has scheduled the generation of non-manual information after the machine translation task and treated it as a sequence classification task. This is justified by the fact that the boundaries of linguistic non-manual components align with those of manual components, rendering the process of generating non-manual components a gloss-labeling task.

Sequence classification is a technique commonly used in the automatic processing of spoken languages. The work we have reported on in this paper is presumably the first to apply it to the generation of non-manual information in sign languages.

We have shown that all of our experimental approaches outperformed their lower baselines. The experimental approaches consisted of: predicting head and eyebrow information together in one label, predicting head and eyebrow information separately, predicting head information by using eyebrow information as additional evidence and vice versa (cascaded approach), and applying an `IOB` format. We have demonstrated the potential of applying a cascaded approach, and we did not observe any benefit from utilizing an `IOB` format in our task.

As a next step, we intend to apply our experimental approaches to a parallel English/American Sign Language (ASL) corpus from a domain with greater semantic and syntactic variety [28]. In this way, we hope to determine, for a different sign language and for a more diverse corpus, if the key findings of this paper are replicable, namely: (a) that the non-manual components of a sign language sentence may be successfully predicted using a sequence classification approach and (b) that some orderings of cascading the sequential predictions are more successful than others (e.g., for DSGS, we found that head information was useful to consider when predicting eyebrow information).

In contrast to the train announcement DSGS corpus used in this paper, the ASL corpus is known to contain instances of non-manual information that are not lexically-cued, i.e., they are not recoverable from the glosses alone. For example, a *yes/no* question in ASL can have the same surface form (gloss order) as a declarative sentence.[5] Thus, we anticipate investigating how to best exploit information contained in the (English) source sentence, e.g., to include question marks as absolute features. Leveraging information from the source sentence will also allow us to capture instances in which a syntactic function is expressed non-manually only on the ASL side: e.g., in ASL, it is possible to convey negation solely via headshake, without the use of any manual sign to indicate the negation [29].

5. References

[1] H. Kacorri, P. Lu, and M. Huenerfauth, "Effect of Displaying Human Videos During an Evaluation Study of American Sign Language Animation," *ACM Transactions on Accessible Computing*, vol. 5, 2013.

[2] O. Crasborn, "Nonmanual Structures in Sign Language," in *Encyclopedia of Language & Linguistics*, 2nd ed., K. Brown, Ed. Oxford: Elsevier, 2006, vol. 8, pp. 668–672.

[3] C. Steiner, *Über die Funktion des Anhebens der Augenbrauen in der Deutschschweizerischen Gebärdensprache DSGS*, ser. Informationsheft. Zürich: Verein zur Unterstützung des Forschungszentrums für Gebärdensprache, 2000, vol. 35.

[4] P. Boyes Braem, *Einführung in die Gebärdensprache und ihre Erforschung*, 3rd ed., ser. Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser. Hamburg: Signum, 1995, vol. 11.

[5] D. Stein, C. Schmidt, and H. Ney, "Analysis, preparation, and optimization of statistical sign language machine translation," *Machine Translation*, vol. 26, pp. 325–357, 2012, published online: 18 March 2012.

[6] S. Morrissey, "Data-Driven Machine Translation for Sign Languages," Ph.D. dissertation, Dublin City University, Dublin, Ireland, 2008.

[7] G. Massó and T. Badia, "Dealing with Sign Language Morphemes in Statistical Machine Translation," in *LREC 2010: 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, La Valetta, Malta, 2010, pp. 154–157.

[8] M. Huenerfauth, "Generating American Sign Language Classifier Predicates for English-to-ASL Machine Translation," Ph.D. dissertation, University of Pennsylvania, 2006.

[9] M. Filhol, "Combining two synchronisation methods in a linguistic model to describe sign language," in *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, ser. LNCS/LNAI. Springer, 2012, vol. 7206.

[10] G. Morgan and B. Woll, Eds., *Directions in sign language acquisition – Trends in language acquisition research*. Amsterdam: John Benjamins Publishing Company, 2002.

[11] V. Jennings, R. Elliott, R. Kennaway, and J. Glauert, "Requirements For A Signing Avatar," in *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010, pp. 133–136.

[12] C. Baker and C. Padden, "Focusing on the non-manual components of ASL," in *Understanding language through sign language research*, P. Siple, Ed. New York: Academic Press, 1978, pp. 27–57.

[13] G. Coulter, "Raised eyebrows and wrinkled noses: The grammatical function of facial expression in relative clauses and related constructions," in *ASL in a bilingual, bicultural context. Proceedings of the Second National Symposium on Sign Language Research and Teaching*, F. Caccamise and D. Hicks, Eds. Coronado, CA: National Association of the Deaf, 1978, pp. 65–74.

[14] S. Liddell, "Non-manual signals and relative clauses in American Sign Language," in *Understanding language through sign language research*, P. Siple, Ed. New York: Academic Press, 1978, pp. 59–90.

[15] ——, *American Sign Language syntax*. The Hague: Mouton, 1980.

[16] C. Baker-Shenk, "A Micro-Analysis of the Nonmanual Components of Questions in American Sign Language," Ph.D. dissertation, University of California, Berkeley, 1983.

[17] J.S. Reilly and M. McIntire and U. Bellugi, "The acquisition of conditionals in American Sign Language: Grammaticized facial expressions," *Applied Psycholinguistics*, vol. 11, pp. 369–392, 1990.

[18] J. Reilly and U. Bellugi, "Competition on the face: Affect and language in ASL motherese," *Journal of Child Language*, vol. 23, pp. 219–239, 1996.

[19] D. Anderson and J. Reilly, "PAH! the acquisition of adverbials in ASL," *Sign Language & Linguistics*, vol. 1, pp. 3–28, 1998.

[20] R. Wilbur, "Phonological and Prosodic Layering of Nonmanuals in American Sign Language," in *The Signs of Language Revisited*, K. Emmorey and H. Lane, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2000, pp. 215–244.

[21] J. Reilly and D. Anderson, "Faces: The acquisition of non-manual morphology in ASL," in *Directions in Sign Language Acquisition*, G. Morgan and B. Woll, Eds. Amsterdam: John Benjamins, 2002, pp. 159–181.

[22] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4, pp. 267–373, 2012.

[23] T. Lavergne, O. Cappé, and F. Yvon, "Practical Very Large Scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, July 2010, pp. 504–513.

[5] The same is true for DSGS, but such cases do not appear in the train announcement data used for the experiments reported in this paper.

[24] E. F. T. K. Sang and J. Veenstra, “Representing Text Chunks,” in *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, ser. EACL ’99. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999, pp. 173–179.

[25] L. Roth and S. Clematide, “Tagging Complex Non-Verbal German Chunks with Conditional Random Fields,” in *Proceedings of the 12th Konvens*, Hildesheim, Germany, 2014, pp. 48–57.

[26] D.-K. Kang, A. Silvescu, and V. Honavar, “RNBL-MN: A Recursive Naive Bayes Learner for Sequences Classification,” in *PAKDD’06*, 2006.

[27] A. Savkov, J. Carroll, and J. Cassell, “Chunking Clinical Text Containing Non-Canonical Language,” in *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014)*, Baltimore, Maryland USA, 2014, pp. 77–82.

[28] C. Neidle and C. Vogler, “A new web interface to facilitate access to corpora: Development of the asllrp data access interface,” in *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012*, Istanbul, Turkey, 2012.

[29] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R. Lee, *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*, 2nd ed. Cambridge: MIT Press, 2001.

Synthesizing the finger alphabet of Swiss German Sign Language and evaluating the comprehensibility of the resulting animations

Sarah Ebling[1], *Rosalee Wolfe*[2], *Jerry Schnepp*[3], *Souad Baowidan*[2],
John McDonald[2], *Robyn Moncrief*[2], *Sandra Sidler-Miserez*[1], *Katja Tissi*[1]

[1]University of Zurich, Zurich, Switzerland
[2]DePaul University, Chicago, IL, USA
[3]Bowling Green State University, Bowling Green, OH, USA
ebling@cl.uzh.ch, {wolfe,jmcdonald}@cs.depaul.edu, schnepp@bgsu.edu,
{rkelley5,sbaowida}@mail.depaul.edu, sandysidler@gmail.com, katja.tissi@hfh.ch

Abstract

This paper reports on work in synthesizing the finger alphabet of Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS) as a first step towards a fingerspelling learning tool for this language. Sign language synthesis is an instance of automatic sign language processing, which in turn forms part of natural language processing (NLP). The contribution of this paper is twofold: Firstly, the process of creating a set of hand postures and transitions for the DSGS finger alphabet is explained, and secondly, the results of a study assessing the comprehensibility of the resulting animations are reported. The comprehension rate of the signing avatar was highly satisfactory at 90.06%.

1. Introduction

Sign languages are natural languages and, as such, fully developed linguistic systems. They are often the preferred means of communication of Deaf[1] signers.

Sign languages make use of a communication form known as the *finger alphabet* (or, *manual alphabet*), in which the letters of a spoken language[2] word are fingerspelled, i.e., dedicated signs are used for each letter of the word. The letters of the alphabet of the most closely corresponding spoken language are used, e.g., English for American, British, and Irish Sign Language; German for German, Austrian, and Swiss German Sign Language, etc. Figure 1 shows the manual alphabet of Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS). Some fingerspelling signs are iconic, i.e., their meaning becomes obvious from their form. Most manual alphabets, like the one for DSGS, are one-handed, an exception being the two-handed alphabet for British Sign Language.

Tools for learning the finger alphabet of a sign language typically display one still image for each letter, thus not accounting for all of the salient information inherent in fingerspelling [3]: According to Wilcox [4], the transitions are more important than the holds for perceiving a fingerspelling sequence. The transitions are usually not represented in sequences of still images.

More recently, 3D animation has been used in fingerspelling learning tools. This approach "has the flexibility to shuffle letters to create new words, as well as having the potential for producing the natural transitions between letters" [3]. The difference between an animation and a still-only representation is shown in Figure 2 for the example of the American Sign Language (ASL) fingerspelling sequence T-U-N-A [5].

This paper reports on the work in synthesizing the finger alphabet of DSGS as a first step towards a fingerspelling learning tool for this language. Sign language synthesis is an instance of automatic sign language processing, which in turn forms part of natural language processing (NLP) [6]. The contribution of this paper is twofold: Firstly, the process of creating a set of hand postures and transitions for the DSGS finger alphabet is explained, and secondly, the results of a study assessing the comprehensibility of the resulting animations are reported. The comprehension rate of the signing avatar was highly satisfactory at 90.06%.

The remainder of this paper is organized as follows: Section 2 gives an overview of previous work involving linguistic analysis (Sections 2.1 to 2.3) and synthesis (Section 2.4) of fingerspelling. Section 3 explains how we produced a set of hand postures and transitions for DSGS fingerspelling synthesis. Section 4 presents the results of the study assessing the comprehensibility of synthesized DSGS fingerspelling sequences.

2. Fingerspelling

2.1. Domains of use

Fingerspelling is often used to express concepts for which no lexical sign exists in a sign language. Apart from that, it may serve other purposes: In ASL, fingerspelling is sometimes applied as a contrastive device to distinguish between "the everyday, familiar, and intimate vocabulary of signs, and the distant, foreign, and scientific vocabulary of words of English origin" [7]. Fingerspelling is also used for quoting from written texts, such as the Bible. In Italian Sign Language, fingerspelling is used predominantly for words from languages other than Italian [7].

Padden and Gunsauls [7], looking at 2164 fingerspelled words signed by 14 native ASL signers, found that nouns are by far the most commonly fingerspelled parts of speech, followed by adjectives and verbs. Within the noun category, occurrences of fingerspelling were evenly distributed among proper nouns and common nouns.

[1] It is a widely recognized convention to use the upper-cased word *Deaf* for describing members of the linguistic community of sign language users and the lower-cased word *deaf* when referring to the audiological state of a hearing loss [1].

[2] *Spoken language* refers to a language that is not signed, whether it be represented in spoken or written form.

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 10–16,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

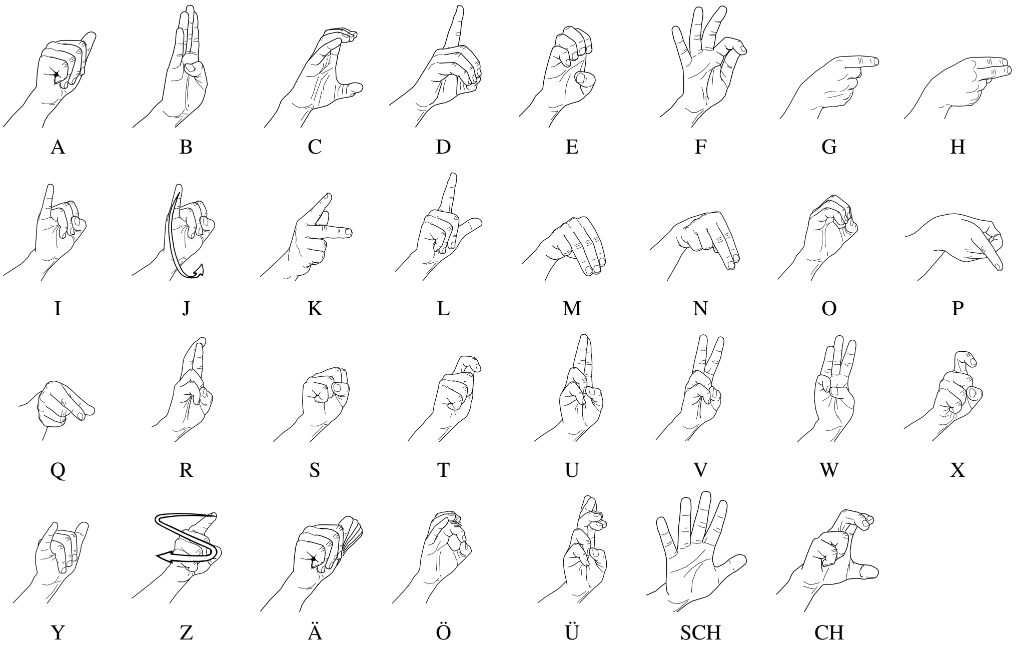


Figure 1: Finger alphabet of DSGS [2]

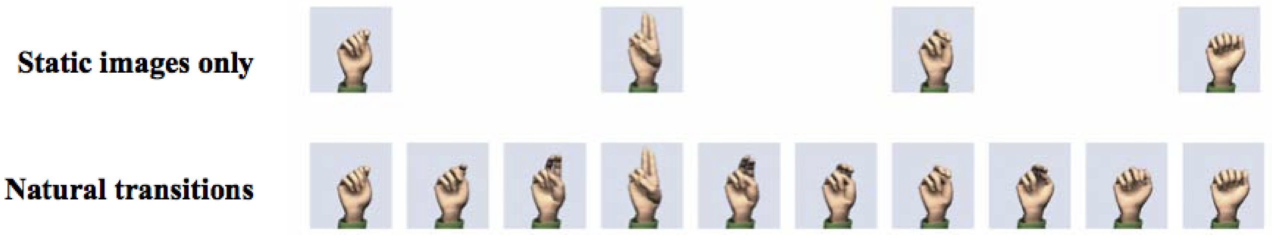


Figure 2: Still images vs. animation: fingerspelling sequence T-U-N-A in American Sign Language [5]

2.2. Frequency of use and speed

Frequency of use and speed of fingerspelling vary across sign languages. ASL is known to make heavy use of the finger alphabet: 10 to 15% of ASL signing consists of fingerspelling [7]. Native signers have been shown to fingerspell more often (18% of the signs in a sequence of 150 signs) than non-native signers (15% of the signs). Within the first group, native signers with a more advanced formal education (college or postgraduate level) have been demonstrated to use more fingerspelling (21% of the signs in a sequence of 150 signs) than native signers at the high school level (15% of the signs) [7].

In ASL, fingerspelled words continue to be used even after lexical signs have been introduced for the same concepts [7]. Some fingerspelled words have also been lexicalized in this language: For example, the sign FAX is performed by signing -F- and -X- in the direction from the subject to the object. This is different from the fingerspelled word F-A-X, which is not reduced to two fingerspelled letters and does not exhibit directionality [7].

Compared to 10 to 15% in ASL, British Sign Language (BSL) has been shown to contain only about 5% fingerspelling [8]. In BSL, fingerspelled words are typically abandoned once lexicalized signs have been introduced for a concept.

In DSGS, fingerspelling is even less common than in BSL. As Boyes Braem and Rathmann [9] pointed out, "few DSGS signers are as yet as fluent in producing or reading fingerspelling".[3] Until recently, DSGS signers used mouthings to express technical terms or proper names for which no lexical sign existed, which partly accounts for the heavy use of mouthing in this language [11].[4] Nowadays, fingerspelling is used more often in these cases, particularly by younger DSGS signers. In addition, it is applied with abbreviations.

Keane and Brentari [13] reported fingerspelling rates between 2.18 and 6.5 letters per second (with a mean of 5.36 letters per second) based on data from different studies. The speed of ASL fingerspelling is known to be particularly high [7], whereas fingerspelling in DSGS is much slower: Accordingly, in a recent focus group study aimed at evaluating a DSGS signing avatar, the seven participants, all of them native signers of DSGS, found the default speed of fingerspelling of the avatar system to be too high [14].

[3] This observation is repeated in Boyes Braem et al. [10].

[4] According to Boyes Braem [12], 80 to 90% of signs in DSGS are accompanied by a mouthing.

2.3. Comprehensibility

A few studies have looked at the comprehensibility of fingerspelling sequences produced by human signers. Among them is that of Hanson [15], who presented 17 Deaf adult signers (15 of which were native signers) with 30 fingerspelled words and non-words each. The participants were given ten seconds to write the letters of the item presented and decide whether it was a word or a non-word.

Geer and Keane [16] assessed the respective importance of holds and transitions for fingerspelling perception. 16 L2 learners of ASL saw 94 fingerspelled words. Each word was presented exactly twice. Following this, the participants were asked to type its letters on a computer. The findings of the study complement those of Wilcox [4] introduced in Section 1: Ironically, the motion between the letters, which is what experts utilize [4], confuses language learners. It is therefore imperative that study tools help language learners learn to decode motion.

2.4. Synthesis

There are three essential elements required for realistic fingerspelling synthesis. These are

- *Natural thumb motion.* Early efforts relied on related work in the field of robotics, however, this proved inadequate as an approximation of the thumb used in many grasping models does not accurately reflect the motions of the human thumb [17].
- *Highly realistically modelled hand with a skeletal deformation system.* Early systems used a segmented hand comprised of rigid components, and lacked the webbing between thumb and index finger, and the ability to deform the palm.
- *Collision detection or collision avoidance.* There is no physicality to a 3D model, so there is no inherent method to prevent one finger from passing through another. Collision detection or avoidance systems can prevent these types of intersections and add to the realism of the model.

An early effort used VRML [18] to allow users to create the hand postures representing individual letters of a manual alphabet. Users could type text and see a segmented hand interpolate between subsequent hand postures. All of the joint coordinates were aligned with world coordinates and did not reflect the natural anatomy of the hand. There were no allowances for collision detection or avoidance.

McDonald [19] created an improved hand model that not only facilitated thumb behavior, but for all of the phalanges in the hand. This was coupled with Davidson's [20] initial work on collision avoidance to produce a set of six words which were tested by Deaf high school students. Although they had few problems in identifying the words, test participants found the appearance of the hand off-putting because it was segmented and lacked webbing between the thumb and index finger.

Adamo-Villani and Beni [21] solved this problem by creating a highly realistic hand model with a skeletal deformation system, allowing the webbing to stretch and wrinkle as does a human hand. In 2006, Wolfe et al. [5] integrated the natural thumb movement and a highly realistic hand model with an enhanced system of collision avoidance. The collision system involved an exhaustive search of all possible letter transitions and correcting any that generated collisions through manual animation.

In 2008, Adamo-Villani [22] confirmed that manually-created animations for fingerspelling are more "readable" than ones generated through motion capture. The research described in this section focused exclusively on ASL, but several groups have explored animating manual alphabets for other signed languages. In 2003, Yeates [23] created a fingerspelling system for Auslan (Australian Sign Language) that utilized a segmented hand; similarly van Zijl [24] and Krastev [25] generated fingerspelling using the International Sign Alphabet. In addition, Kennaway [26] explored fingerspelling for BSL.

While only a small body of work has dealt with the comprehensibility of fingerspelling produced by human signers, even fewer studies have investigated the comprehensibility of synthesized fingerspelling. Among them is the study of Davidson et al. [20], who presented fluent ASL users with animated fingerspelling sequences at three different speeds to validate their animation approach.

3. Creating a set of hand postures and transitions for DSGS fingerspelling synthesis

Section 2.2 discussed the increasing use of fingerspelling in DSGS. To our knowledge, only one fingerspelling learning tool for DSGS exists.[5] This tool displays one illustration for each letter of a fingerspelling sequence as mentioned in Section 1. Ours is the first approach to synthesizing the finger alphabet of DSGS as a first step towards a learning tool for this language.

Synthesizing the DSGS manual alphabet consisted of producing hand postures (handshapes with orientations) for each letter of the alphabet and transitions for each pair of letters. Figure 1 showed the finger alphabet of DSGS. Note that it features dedicated signs for -Ä-, -Ö-, and -Ü- as well as for -CH- and -SCH-.

Because of the similarity between the ASL and DSGS manual alphabets, our work built on a previous system that synthesized the manual alphabet of ASL [5]. In addition to the five new letters or letter combinations cited above, the DSGS manual alphabet contains four handshapes, -F-, -G-, -P-, and -T-, that are distinctly different from ASL. Further, the five letters -C-, -M-, -N-, -O-, and -Q- have a similar handshape in DSGS, but required smaller modifications, such as a different orientation or small adjustments in the fingers. Hence, the DSGS finger alphabet features 14 out of 30 hand postures that needed modification from the ASL manual alphabet. All hand postures were reviewed by native signers.

Like ASL, there was also the issue of collisions between the fingers during handshape transitions. Here, we again leveraged the similarity between ASL and DSGS manual alphabets. The previous ASL fingerspelling system identified the collection of letter pairs, such as the N→A transition in T-U-N-A in Figure 2, which caused finger collisions under naïve interpolation. To remove the collisions, they created a set of transition handshapes that are inserted in-between two letters to force certain fingers to move before others to create the clearance needed to avoid collision. Such a handshape can be seen in the eighth frame of the second row in Figure 2. Details of this method can be found in Wolfe et al. [5]. Because of the overlap between the DSGS and ASL manual alphabets, along with the fact that most of the new or modified hand postures had handshapes that were generally open, in the sense of Brentari's hanshape notation [27], it

[5] http://www.gebaerden-sprache.ch/fingeralphabet/lernen-sie-das-fingeralphabet/index.html

was possible to use the exact same set of transition handshapes as the original ASL system.

4. Assessing the comprehensibility of synthesized DSGS fingerspelling sequences

The aim of the study presented here was to assess the comprehensibility of animated DSGS fingerspelling sequences produced from the set of hand postures and transitions described in Section 3.

4.1. Study instrument and design

We conducted the study online using a remote testing system, *LimeSurvey*[6]. This approach has advantages over to face-to-face testing because it affords a large recruitment area and allows participants to complete the survey at any time. The survey was accessible from most web browsers and compatible across major operating systems.

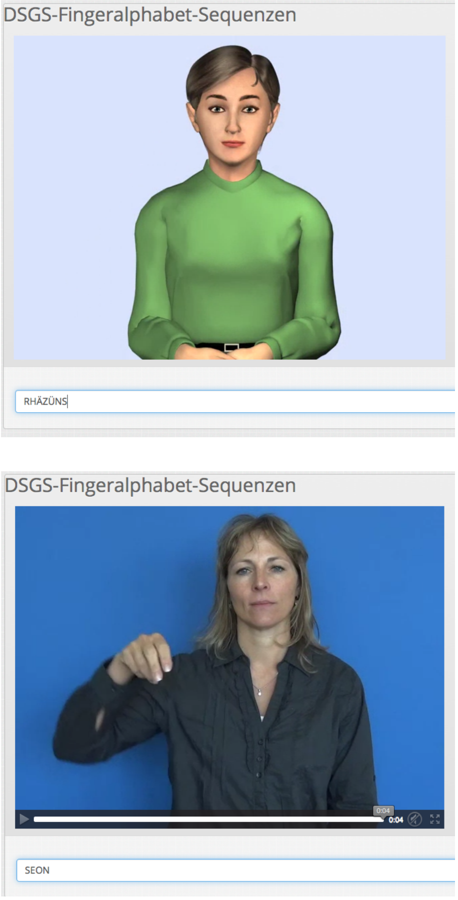
Any person with DSGS fingerspelling skills was invited to participate in the study. The call for participation was distributed via an online portal for the DSGS community[7] as well as through personal messages to persons known to fulfill the recruitment criteria.

Participants accessed the study through a URL provided to them. The first page of the website presented information about the study in DSGS (video of a human signer) and German (video captions that represented a back-translation of the DSGS signing and text). Participants were informed of the purpose of the study, that participation was voluntary, that answers were anonymous, that items could be skipped, and that they could fully withdraw from the study at any time. Following this, they filled out a background questionnaire, which included questions about their hearing status, first language, preferred language, and age and manner of DSGS acquisition. No personally identifyable information was kept.

A detailed instruction page followed, on which the participants were informed that they were about to see 22 fingerspelled words signed by either a human or a virtual human (sign language avatar). Following this, the participants' task was to type the letters of the word in a text box. Figure 3 shows a screenshot of the study interface for each of these cases. The videos of the human signer had been resized and cropped so as to match the animations.

The participants were told that the fingerspelled words they were going to see were names of Swiss towns described in Ebling [14]. In contrast to the studies discussed in Section 2.3, an effort had been made to include only fingerspelled words that denote concepts for which no well-known lexical sign exists in DSGS. This was deemed an important prerequisite for a successful study. The items had been chosen based on the following criteria:

- They were names of towns with train stations that were among the least frequented based on a list obtained from the Swiss Federal Railways;
- The town names were of German or Swiss German origin;
- The town names in the resulting set of items varied with respect to their length (number of letters); and
- In the resulting set of items, each letter of the DSGS finger alphabet occurred at least once (with the exception of -X-, which did not occur in any of the town names that met all of the above criteria).



DSGS-Fingeralphabet-Sequenzen

RHÄZÜNS

DSGS-Fingeralphabet-Sequenzen

SEON

Figure 3: Study interface: screenshots

The 20 study items had an average length of 7 letters, with a maximum of 12 (W-E-R-T-H-E-N-S-T-E-I-N) and a minimum of 3 (T-Ä-SCH). The study items were assigned to participants such that each item appeared as either a video of a human signer or as an animation. Each participant saw 10 videos and 10 animations and items were presented in random order. The study items were preceded by two practice items that were the same for all participants: The first was a video of a human signer fingerspelling S-E-O-N, the second an animation of R-H-Ä-Z-Ü-N-S.

The human signer was a female native DSGS signer (Deaf-of-Deaf) who had been asked to sign at a natural speed but without using mouthings. This resulted in an average fingerspelling rate of 1.76 letters per second. The same rate was used for the animations. Note that it is below the minimum rate of 2.18 reported by Keane and Brentari [13] (cf. Section 2.2), which again points in the direction of a lower speed of fingerspelling in DSGS.

The participants were informed that they could view a video as many times as they wanted. Limiting the number of viewings

[6] https://www.limesurvey.org/en/

[7] http://www.deafzone.ch/

was felt to exert undue pressure. This approach was different from the study of Geer and Keane [16] (Section 2.3), who allowed subjects to view a video exactly twice, and Hanson [15], who presumably showed each video once. Not restricting the number of viewings in the present study also meant that there was no limit to the response time for an item. The response time was recorded as metadata.

Once participants had completed the main part of the study, they were asked to provide feedback on the following aspects:

- Appropriateness of the rate of fingerspelling;
- Comprehensibility of the individual letters and transitions between letters; and
- General feedback on the fingerspelling sequences shown

On the final page, participants were thanked for their contribution and given the possibility to leave their e-mail address if they wanted to receive information on the results of the study. If provided, the e-mail address was not saved together with the rest of the data to ensure anonymity. All data was stored in a password-protected database.

The entire study was designed so as to take a maximum of 20 minutes to complete. This was assessed through a pilot study with three participants, in which the average time spent to complete the study was 17 minutes.

4.2. Results and discussion

The study remained online for one week. During this time, 65 participants completed it, of which 31 were hearing, 24 Deaf, and 6 hard-of-hearing. 4 participants indicated that they did not fall into the three categories proposed for hearing status, referring to themselves as "using sign and spoken language", "deafened", "CODA" (child of Deaf adult), and "residual hearing/profoundly hard-of-hearing". The average time taken to complete the entire survey was 20 minutes and 12 seconds.

For the 20 main study items (excluding the two practice items), 1284 responses were submitted. In relation to the 1300 possible responses (20 items × 65 participants), this meant that a total of 16 responses had been skipped.[8] They were treated as incorrect responses.

For each of the 1284 responses given, we determined whether it was correct, ignoring umlaut expansions (*ä→ae*, etc.) and differences in case. Table 1 displays the comprehension rates: The mean percentage of correct responses was 93.91% for sequences fingerspelled by the human signer and 90.06% for sequences fingerspelled by the avatar. Also displayed are the binomial confidence intervals at a confidence level of 95%. They indicate a 95% confidence that the comprehension rate of the signing avatar is above 87.75% and below 92.37%. This result is highly satisfactory.

Comprehension rates below 100% for human signing have been reported in previous studies [28, 29]. We surmise that in this case, they were due at least partly to the fact that mouthings were absent from the signing performances. While this was a methodological decision made to ensure that what was being measured was core fingerspelling comprehension, several participants alluded to the lack of mouthings in the post-study questionnaire.

A comprehension rate of 100% was obtained for three sequences fingerspelled by the human signer (*Realp*, *Reutlingen*, and *Sedrun*) and also for three sequences produced by the signing avatar (*Bever*, *Hurden*, and *Mosen*).

To obtain information about individual letters that may have been hard to comprehend with the signing avatar, we performed a confusion analysis. The results show that three letters were mistaken for other letters more often in sequences fingerspelled by the signing avatar than in sequences fingerspelled by the human signer: -F- (confused with -T- and -B-), -P- (confused with -G- and -H-), and -R- (confused with -U-). One letter, -H-, was confused more often in sequences fingerspelled by the human signer than in sequences fingerspelled by the signing avatar; it was mistaken with -G-, -L-, and -U-.

A confusion analysis between pairs of letters was also performed to obtain pointers to transitions that potentially needed to be improved. Comprehension was lower for four transitions with the signing avatar than with the human signer: F-I (mistaken for T-I and B-I), L-P (mistaken for L-G and L-H), L-R (mistaken for L-U), and R-I (mistaken for U-I). This overlaps with the qualitative feedback in the post-study questionnaire that asked for letters and transitions that were particularly hard to understand: Several participants mentioned the avatar's transitions into -G-, -I-, -P-, and -Q- as well as the transitions between -D- and -Q- and -L- and -P-. In addition, 12 out of 65 participants deemed the hand orientation of -Q- inaccurate.

In the general comments section, a number of participants remarked that the fingerspelling of the human signer was easier to understand than that of the signing avatar; some participants noted that this was due to the hand appearing too small in the animations. On the other hand, multiple participants commented on the quality of the signing avatar as being "surprisingly good". Repeated mention was made of the impression that short fingerspelled sequences were easier to understand than longer ones, regardless of whether they were signed by a human or an avatar.

One participant encouraged the introduction of speed controls for the signing avatar. In the post-study questionnaire rating of the speed of fingerspelling, the majority of the participants (number of responses: 62) deemed the speed appropriate (56.45%), followed by 35.48% who rated it as being too fast. 4.84% classified it as too slow, and 3.23% deemed it much too fast. No participant rated the speed as being much too slow. The numbers are summarized in Table 2.

5. Conclusion and outlook

We have presented the first work in synthesizing the finger alphabet of DSGS, an application of natural language processing. We have reported on the process of creating a set of hand postures and transitions as well as on a study assessing the comprehensibility of the resulting animations. The results showed that the comprehension rate of the signing avatar was highly satisfactory at 90.06%. Three of the sequences fingerspelled by the avatar yielded a comprehension rate of 100%.

The speed of fingerspelling chosen for the signing avatar was rated as appropriate by the majority of the participants. At the same time, a lower yet substantial number of participants rated it as being too high, which suggests that introducing speed controls would be beneficial.

The results of the study also offered pointers to aspects of the signing avatar that would benefit from further improvement, such as the hand postures of a number of letters as well as the transitions between some letters.

While the primary aim of the study was to assess the comprehensibility of the newly-created DSGS fingerspelling animations, the data obtained provides a wealth of information that

[8] Recall that participants were given the option of not responding at any point in the study.

	Comprehension rate (%)	**Confidence interval lower bound (%)**	**Confidence interval upper bound (%)**
Human signer	93.91	92.05	95.76
Signing avatar	90.06	87.75	92.37

Table 1: Percentage of correct responses

Rating	**Responses (%)**
much too slow	0.00
too slow	4.84
appropriate	56.45
too fast	35.48
much too fast	3.23

Table 2: Speed of fingerspelling: rating

can be used to inform other research questions. For example, we intend to investigate the individual effects of the variables hearing status, age of DSGS acquisition, and speed-of-fingerspelling rating on the comprehension scores.

The work presented in this paper represents the first step towards a fingerspelling learning tool for DSGS. As a next step, we will complete the development of the tool interface. Following this, we are going to conduct a study that assesses the usability of the interface.

6. References

[1] G. Morgan and B. Woll, Eds., *Directions in sign language acquisition – Trends in language acquisition research*. Amsterdam: John Benjamins Publishing Company, 2002.

[2] P. Boyes Braem, "A multimedia bilingual database for the lexicon of Swiss German Sign Language," *Sign Language & Linguistics*, vol. 4, pp. 133–143, 2001.

[3] J. Toro, J. McDonald, and R. Wolfe, "Fostering Better Deaf/Hearing Communication through a Novel Mobile App for Fingerspelling," in *ICCHP 2014, Part II, LNCS 8548*, K. M. et al., Ed., 2014, pp. 559–564.

[4] S. Wilcox, *The Phonetics of Fingerspelling*. Amsterdam: John Benjamins Publishing, 1992.

[5] R. Wolfe, N. Alba, S. Billups, M. Davidson, C. Dwyer, D. G. Jamrozik, L. Smallwood, K. Alkoby, L. Carhart, D. Hinkle, A. Hitt, B. Kirchman, G. Lancaster, J. McDonald, L. Semler, J. Schnepp, B. Shiver, A. Suh, and J. Young, "An Improved Tool for Practicing Fingerspelling Recognition," in *Conference 2006 International Conference on Technology and Persons with Disabilities*, Northridge, California, March 2006, pp. 17–22.

[6] E. Sáfár and J. Glauert, "Computer modelling," in *Sign Language. An International Handbook*, R. Pfau, M. Steinbach, and B. Woll, Eds. Berlin, Boston: De Gruyter Mouton, 2012, ch. 44, pp. 1075–1101.

[7] C. Padden and D. Gunsauls, "How the Alphabet Came to Be Used in a Sign Language," *Sign Language Studies*, vol. 4, pp. 10–33, 2003.

[8] M. Brennan, "Making Borrowing Work in British Sign Language," in *Foreign Vocabulary in Sign Languages: A Cross-Linguistic Investigation of Word Formation*, D. Brentari, Ed. Mahwah, N.J.: Erlbaum, 2001, pp. 49–85.

[9] P. Boyes Braem and C. Rathmann, "Transmission of sign language in Northern Europe," in *Sign languages*, ser. Cambridge language surveys, D. Brentari, Ed. Cambridge: Cambridge University Press, 2010, pp. 19–45.

[10] P. Boyes Braem, T. Haug, and P. Shores, "Gebärdenspracharbeit in der Schweiz: Rückblick und Ausblick," *Das Zeichen*, vol. 90, pp. 58–74, 2012.

[11] P. Boyes Braem, *Einführung in die Gebärdensprache und ihre Erforschung*, 3rd ed., ser. Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser. Hamburg: Signum, 1995, vol. 11.

[12] ——, "Functions of the Mouthing Component in the Signing of Deaf Early and Late Learners of Swiss German Sign Language," in *Foreign Vocabulary in Sign Languages: A Cross-Linguistic Investigation of Word Formation*, D. Brentari, Ed. Mahwah, NJ: Erlbaum, 2001, pp. 1–47.

[13] J. Keane and D. Brentari, *The Oxford Handbook of Deaf Studies in Language: Research, Policy, and Practice*. Oxford University Press, forthcoming, ch. Fingerspelling: Beyond Handshape Sequences.

[14] S. Ebling, "Evaluating a Swiss German Sign Language Avatar among the Deaf Community," in *Proceedings of the Third International Symposium on Sign Language Translation and Avatar Technology*, Chicago, IL, 2013.

[15] V. Hanson, "When a Word Is Not the Sum of Its Letters: Fingerspelling and Spelling," in *Teaching American Sign Language as a Second/Foreign Language: Proceedings of the Third National Symposium on Sign Language Research and Teaching*, F. Caccamise, M. Garretson, and U. Bellugi, Eds., 1981, pp. 176–185.

[16] L. Geer and J. Keane, "Exploring Factors that Contribute to Successful Fingerspelling Comprehension," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[17] M. Girard, "Constrained optimization of articulated animal movement in computer animation," in *Making Them Move*. Morgan Kaufmann Publishers Inc., 1991, pp. 209–232.

[18] S. A. Su and R. K. Furuta, "VRML-based representations of ASL fingerspelling on the World Wide Web," in *Proceedings of the third international ACM conference on Assistive technologies*, 1998, pp. 43–45.

[19] J. McDonald, J. Toro, K. Alkoby, A. Berthiaume, R. Carter, P. Chomwong, and J. C. et al., "An improved articulated model of the human hand," *The Visual Computer*, vol. 17, pp. 158–166, 2001.

[20] M. J. Davidson, K. Alkoby, E. Sedgwick, R. Carter, J. Christopher, B. Craft, and J. F. et al., "Improved Hand Animation for American Sign Language," in *Technology And Persons With Disabilities Conference*, 2001.

[21] N. Adamo-Villani and G. Beni, "Automated finger spelling by highly realistic 3d animation," *British journal of educational technology*, vol. 35, pp. 345–362, 2004.

[22] N. Adamo-Villani, "3d rendering of american sign language finger-spelling: a comparative study of two animation techniques," *International Journal of Human and Social Sciences*, vol. 3, p. 24, 2008.

[23] S. Yeates, H. Eun-Jung, and R. Owens, "An animated Auslan tuition system," *Machine Graphics And Vision*, vol. 12, pp. 203–214, 2003.

[24] L. van Zijl and L. Raitt, "Implementation experience with collision avoidance in signing avatars," in *Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*. ACM, 2004, pp. 55–59.

[25] A. Krastev, A. Lekova, M. Dimitrova, and I. Chavdarov, "An interactive technology to support education of children with hearing problems," in *Proceedings of the 15th International Conference on Computer Systems and Technologies*. ACM, 2014, pp. 445–451.

[26] R. Kennaway, "Experience with and Requirements for a Gesture Description Language for Synthetic Animation," in *Gesture-Based Communication in Human-Computer Interaction*, ser. Lecture Notes in Computer Science, A. Camurri and G. Volpe, Eds. Berlin/Heidelberg: Springer, 2004.

[27] D. Brentari, *A Prosodic Model of Sign Language Phonology*. Cambridge, MA: MIT Press, 1998.

[28] M. Kipp, A. Heloir, and Q. Nguyen, "Sign language avatars: Animation and comprehensibility," in *Proceedings of IVA 2011*, ser. Lecture Notes in Computer Science, H. Vilhjálmsson, S. Kopp, S. Marsella, and K. Thórisson, Eds., vol. 6895. Reykjavik, Iceland: Springer Berlin Heidelberg, 2011, pp. 113–126.

[29] WebSourd, "DictaSign Deliverable D7.4: Prototype evaluation synthesis," DictaSign project, Tech. Rep., 2011.

Contour-based Hand Pose Recognition for Sign Language Recognition

Mika Hatano, Shinji Sako, Tadashi Kitamura

Graduate School of Engineering, Nagoya Institute of Technology
{pia, sako, kitamura}@mmsp.nitech.ac.jp

Abstract

We are developing a real-time Japanese sign language recognition system that employs abstract hand motions based on three elements familiar to sign language: hand motion, position, and pose. This study considers the method of hand pose recognition using depth images obtained from the Kinect v2 sensor. We apply the contour-based method proposed by Keogh to hand pose recognition. This method recognizes a contour by means of discriminators generated from contours. We conducted experiments on recognizing 23 hand poses from 400 Japanese sign language words.

Index Terms: hand pose, contour, sign language recognition, real-time, Kinect

1. Introduction

In Japan, Japanese sign language is usually used among hearing impaired people to communicate. In addition, these people often communicate with others through a third person who understands both oral and sign language. The alternative is to use a computer that acts as an interpreter. However, no practical sign language recognition system exists, even one that recognizes isolated words. The difficulties lie in the nature of visual language and its complex structure. Compared with speech recognition, sign language recognition incorporates various visual components, such as hand motions, hand poses and facial expressions. In addition, no established study exists on representing the structure of Japanese sign language in a similar manner to that of spoken language. Therefore, few attempts recognize sign language by units such as hand motions and hand poses [1, 2].

Our study develops with real-time recognition of sign language words. In Japanese sign language, a sentence consists of several words and non-manual signals such as facial expressions. To recognize words is a first step and essential to recognize sentences. The number of Japanese sign language words is said to be 3,000 or more. Recognition by discriminators that are independent of every word has proven ineffective. To produce a practical system, analysis and reconstruction of sign language words are critical. We want to emphasize that database of sign language words is required when we analyze such words. However, no established database currently exists for sign language recognition. Therefore, we employ a database from a computerized sign language word dictionary instead.

Our system is based on three elements of sign language: hand motion, position, and pose. This study considers the method of hand pose recognition for our system. Speeding up hand pose recognition is difficult, because of the number and variety of hand poses caused by rotations, altering the angle from the sensor, and diversities in bone structures. This study considers a hand pose recognition using depth images obtained from a single depth sensor. We apply the contour-based method proposed by Keogh [3] to hand pose recognition. This method recognizes a contour by means of discriminators learned from contours. We conducted experiments to recognize 23 hand poses from 400 Japanese sign language words.

2. System overview

Figure 1 shows the flowchart of the entire system. We use Kinect v2 sensor [4] to obtain data from sign motions produced by an actual person. First, data obtained from the sensor is segmented into sign language words. Second, the three aforementioned elements are recognized individually. Finally, the recognition result is determined by the weighted sum of each score. The recognition process of the hand pose and other two components employs depth data of the hand region and coordinates of joints, respectively. This study partially considers the method of hand pose recognition and does not discuss other processes on the flowchart.

To utilize the structure in sign language recognition requires an expert knowledge of sign language. We apply a database from the computerized sign language word dictionary produced by Kimura [5] to sign language recognition. Our hand pose recognition is based on the classification of hand types employed in this dictionary. Table 1 shows a portion of the database in the dictionary. This database includes approximately 2,600 Japanese sign language words. Each word is represented by specific sign language types in Table 2 and other elements are indicated in Figure 2. For example, the word "red" which belongs to the type 1 in Table 2 is expressed by the dominant hand and the other hand is not used.

3. Method of hand pose recognition

Some methods of hand pose estimation classify depth pixels into parts to obtain joint coordinates [6, 7]. However, these methods present difficulties when the palm does not face the

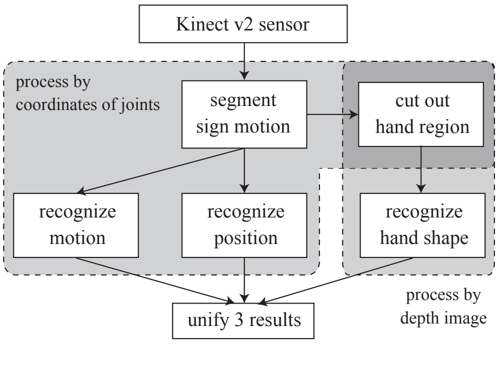


Figure 1: *Flowchart of the entire system.*

Table 1: *Portion of the database in the dictionary.*

Word	SL Type	Hand type	Palm direction	Position	Motion
love	3	B	down	NS	circle
between	4	B	side	NS	down
blue	1	B	back	lower face	back
red	1	1	back	lower face	right
baby	4	B	up	NS	up-down
autumn	4	B	back	whole face	front-back
open	4	B	front	NS	right
morning	1	S	side	upper face	down
shallow	2	B	side	NS	up
tomorrow	1	1	front	whole face	front
play	4	1	side	upper face	front-back
rain	4	5	back	NS	up-down
walk	1	U	back	NS	front
relief	4	B	back	body	down
say	1	1	side	lower face	front

⋮

Table 2: *Sign Language (SL) types.*

	1	2	3	4	5
use both hands	×	○	○	○	○
hand type is same through two hands		○	×	○	×
non-dominant hand moves		×	×	○	○

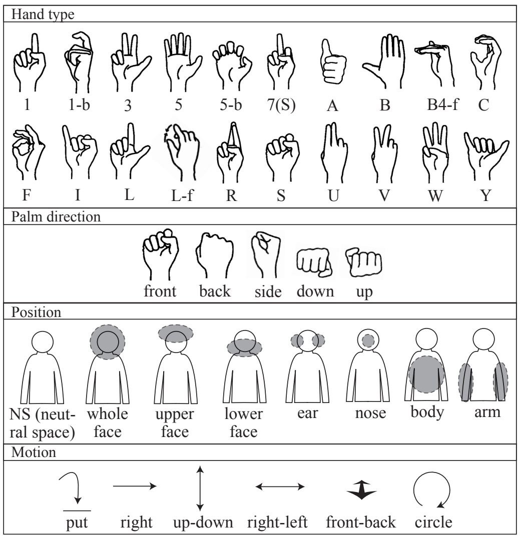


Figure 2: *Elements in sign language dictionary.*

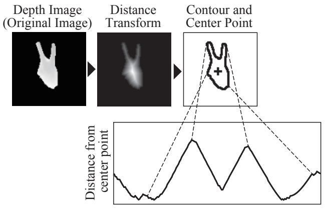


Figure 3: *Feature extraction from an image of a hand region.*

camera and some fingers are invisible. We use the contour-based method proposed by Keogh [3]. Contour-based methods work efficiently when recognition objects have distinct shapes. This method treats a contour that encircles an area as a recognition object and uses discriminators called *wedges* generated from contours. This method is described below.

3.1. Feature extraction

Shapes can be converted to *distance vectors* to form one-dimensional series. Figure 3 shows the procedure for extracting a distance vector from a hand image. First, the center point of the hand region is determined by distance transform. Distance transform labels each pixel whose value is “1” with the distance to the nearest pixel whose value is “0” in a binary image. The center point is a pixel that has a maximal value after distance transform. Next, each distance from the center point to every pixel on the contour is calculated. The distance vector represents a series of these distances.

3.2. Calculation of distance

A distance D between two distance vectors $P = \{p_0, p_1, ..., p_n\}$ and $Q = \{q_0, q_1, ..., q_n\}$ is calculated according to the followings.

$$D(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (1)$$

If the length of two distance vectors is different, dynamic time warping (DTW) should be used to adjust for size variations. However, we do not use DTW to avoid excessive computation time. Instead, we unify their length by resizing them in advance.

We can compare contours by calculating their distances or using discriminators generated from contours. These discriminators are called *wedges*. Wedges have maximal and minimal values at each point. If a contour is located inside a wedge, the distance is 0. The distance D between a wedge W ($U = \{u_0, u_1, ..., u_n\}$ is its top, $L = \{l_0, l_1, ..., l_n\}$ is its bottom) and a contour $P = \{p_0, p_1, ..., p_n\}$ is calculated based on the following equation. For example, the sum of broken lines in Figure 4 is a distance between W and P.

$$D(W,P) = \sqrt{\sum_{i=1}^{n} \begin{cases} (p_i - u_i)^2 & (p_i > u_i) \\ (p_i - l_i)^2 & (p_i < l_i) \\ 0 & (otherwise) \end{cases}} \qquad (2)$$

3.3. Producing wedges

Wedges are produced according to the following procedures.

1. Extract features from hand images.
2. Calculate distances of all contours.
3. Unify two contours in ascending order of distances. The maximal and minimal values of merged contours become a wedge.
4. Repeat process 3. until the number of wedges decreases to a definite number.

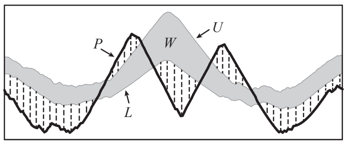


Figure 4: *Distance between a wedge and contour.*

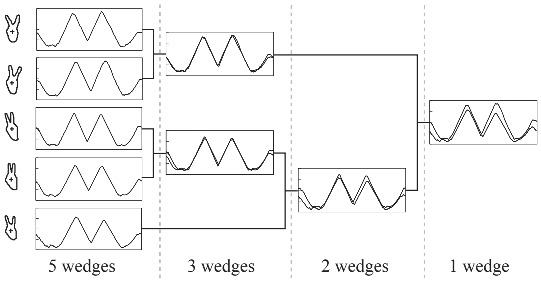


Figure 5: *Producing wedges from five contours.*

When Figure 5 shows an example of producing wedges. A wide wedge produced by contours that are diverse does not function as a discriminator. We prepare various wedges for recognizing each hand type in order to consider the details of contours.

3.4. Speeding up calculation

When we consider a rotation invariant matching of two distance vectors, the calculation must be repeated many times with shifting one of the distance vectors. We can speed up this computation by aborting when the current sum of squared differences exceeds a threshold. In addition, although existing research does not attempt this, we try to speed up the calculation by means of the followings.

- The length of the distance vectors is unified and shortened, and the accuracy does not diminish.
- When the number of wedges per hand type is greater than one, recognition that uses one-by-one wedge is performed prior to help targeting candidates.

4. Experiments

4.1. Datasets

We conducted experiments on recognizing 23 hand poses in 400 Japanese sign language words in the national sign language test grade 5. To recognize these 400 words requires to distinguish 23 hand poses in Table 3 defined by hand types and palm directions. Some words have the same hand poses but different position and motion. Our system distinguish each word after recognizing 3 components and unifying recognition results.

Because hand shapes transform with motions, each hand type remains independent even if the palm direction is different. However, some exceptions exist to distinguish sign language words that have the same motion, position, and hand type, but have a different palm direction. For example, Groups 3 and 4 in Table 3 should be distinguished even though the hand type is the same.

To simplify the collection of data in our experiments, we used depth images of stationary hands instead of those obtained during natural sign motions. Table 4 shows the experimental conditions. We conducted four experiments examining the robustness of the recognition method about the variety of hand shapes and the computation time. The objectives of the experiments are described as follows.

Experiment 1 Recognize 100 hand images by wedges produced from the same 100 images per hand type, palm direction, and tester (close-dataset, close-tester).

Experiment 2 Recognize 50 hand images by wedges produced from the other 50 hand images per hand type, palm direction, and tester (open-dataset, close-tester). Experiments were repeated with different data.

Table 3: *List of 23 hand pose groups.*

ID	Hand type	Palm direction
0	1	front-back, right-left
1	1-b	right-left
2	3	front-back
3	5	front-back
4	5	up-down
5	5-b	front-back, right-left, up-down
6	7(S)	front-back
7	A	front-back, right-left
8	B	front-back
9	B	right-left
10	B	up-down
11	B4-f	right-left
12	C	right-left
13	F	front-back
14	I	front-back
15	L	front-back
16	L-f	right-left
17	R	right-left
18	S	front-back, right-left, up-down
19	U	front-back
20	V	front-back
21	W	front-back
22	Y	front-back

Table 4: *Experimental condition.*

Hand type	20 types in Figure 2
Palm direction	3 patterns (front-back, right-left, up-down)
Hand pose group	23 groups* *determined by hand types and palm directions
Tester's profile	A (female, hand size* 16 cm) B (female, hand size* 18 cm) C (male, hand size* 19 cm) D (male, hand size* 21 cm) *measured from the wrist to the tip of the middle finger
Depth image	100×100 pixel 100 images of the hand region per tester, hand type and palm direction
Length of distance vector	30 or 180
PC specs	OS：Windows 8.1 64 bit RAM: 4 GB CPU：Intel Core i5-4570 (3.20 GHz, 4-core)

Experiment 3 Recognize 100 hand images of a person by wedges produced from 300 hand images of the other three persons per hand type, and palm direction (open-dataset, open-tester). Experiments ware repeated with different data.

Experiment 4 Examine the relationship between the computation time required to recognize a hand image and the average recognition rate from Experiment 2. We attempted to speed up the calculation by the methods in Section 3.4. The threshold value when the calculation was aborted was determined by the preliminary experiment. The length of distance vectors was 30 in this experiment. Each recognition was aided to target candidates as many as five hand pose groups by the recognition that uses one-by-one wedge performed prior.

4.2. Results

4.2.1. Experiment 1, 2

Figure 6 shows the average recognition rates for Experiment 1 and 2. The accuracy can be improved by increasing the number of wedges. This can be accomplished because of the variety of hand shapes caused by posing of hand and by altering the angle from the camera.

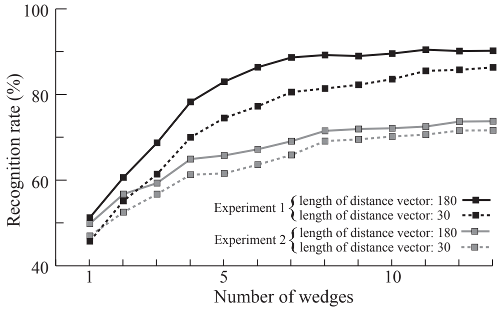


Figure 6: *Experiment 1, 2: Recognition rate and number of wedges per hand type and palm direction (person is known).*

Experiment 1 was conducted with close-dataset. This experiment is just for sanity check and its condition is impossible in real-life. The result was sufficient for our system. Erroneous recognition in Experiment 1 was primarily caused by misidentifying hand pose Groups 4 and 5. These two groups have a common point that includes a hand pose whose palm direction is down. When we obtain data from a single depth camera, capturing the characteristics of hand shapes when the palm does not face the camera is difficult. Group 6 had the lowest recognition rate among the hand pose group (when the length of the distance vector is 180, the number of wedges was 10 per hand type and palm direction, and the group's recognition rate was 80%). This is because the group was misrecognized as Group 0. These two groups have similar shapes. In addition, the recognition rates of Group 2, 13, 15, 20, and 22 were high under all conditions because other groups do not possess similar shapes.

Experiment 2 was conducted with open-dataset and close-tester. The result showed a similar trend to that of Experiment 1 concerning the causes of erroneous recognition. Because no hand shapes from the learning data are included in the evaluation data, the recognition rate was lower than that of Experiment 1. However, no significant difference in recognition rate of Experiment 1 and 2 appeared when the number of wedges is one per hand type and palm direction. Therefore, if the wedges are generated from samples of a certain number, applying unknown data from the same person is possible. The recognition rate from Experiment 2 is expected to approach that of Experiment 1 by increasing the amount of learning data.

Experiments were conducted after changing the length of distance vectors. Although shortening the distance vectors reduces the calculations, the accuracy is expected to fall because of the loss of detailed features. However, no significant differences between the experiments appeared when the length of the distance vectors is 30 and 180. Therefore, if small sized hand images are used or the contours are rough because of noises, a robust recognition can be accomplished.

The maximal number of wedges was between 20 and 25 in Experiment 1 and between 8 and 13 in Experiment 2. The number fluctuated with the complexity of the hand types.

4.2.2. *Experiment 3*

Experiment 3 was tester-independent setup. Figure 7 shows the results of Experiment 3. The recognition rates shown are the results when the length of distance vectors is 30. If we change the length to 180, recognition rates do not change significantly. We specified causes of erroneous recognition when the number of wedges is 30 per hand type and palm direction. The results show the same tendency as in Experiments 1 and 2, that is, 13 % of all data were misrecognized as Groups 4 and 5. The detailed findings for each hand pose group reveal the following: 41 % of Group 6 were misrecognized as Group 0, 53 % of Group 19 were misrecognized as Group 0, 45 % of Group 12 were misrecognized as Group 5.

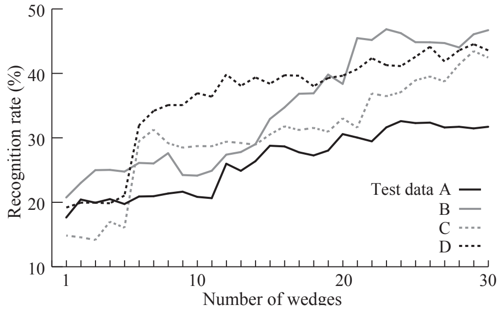


Figure 7: *Experiment 3: Recognition rate and number of wedges per hand type and palm direction (person is unknown).*

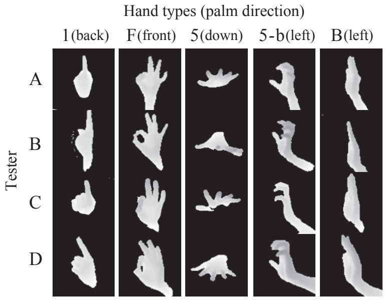


Figure 8: *Variety of hand shapes among people.*

The low recognition rate is due to individual differences in hand shapes caused by differences in bone structure and posing of hand shown in Figure 8. Wedges produced from the hand images of various people include other hand types. This caused misrecognitions.

Per person details show that the recognition rate was lowest when the system attempted to recognize hand poses of tester A, whose hand size was the smallest. When the number of wedges increases, the recognition rate of tester B, whose hand size is between that of A and C is higher than that of other testers.

Although we normalized the scale of distance vectors according to each hand size, hand pose recognition by contours possesses other difficulties when the bone structures are considered. The accuracy diminishes when the system recognizes hand images of a person whose bone structure is dissimilar to any learning data. When we want to recognize hand poses of an unknown person, wedges generated from people who have similar bone structure should be used. Therefore, additional hand images that reveal various characteristics in bone structures should be collected.

4.2.3. *Experiment 4*

Experiment 4 was for checking the computation time. Figure 9 shows the relationship between the computation time required to recognize a hand image and the average recognition rate in Experiment 2. The speed-up process did not affect the recognition rate.

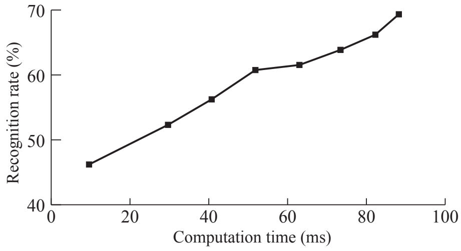


Figure 9: *Experiment 4: Average computation time and the recognition rate required to recognize a hand image.*

When the person is known, 88 ms (corresponding to 11 fps) was required to recognize a hand image with 70 % accuracy. Recognizing all hand images obtained from the sensor with a frame rate of 30 fps is impossible. However, the number of frames required to specify a hand pose is limited because the hand pose does not change at every frame. We can recognize in real-time selected hand images by means of comparison method employing a small calculation such as *image moment* [8]. This experiment has been implemented in a single-thread. The processing speed can be improved by utilizing a high-speed technique such as multi-threading.

5. Conclusion

We are developing a real-time Japanese sign language recognition system based on three elements of sign language: motion, position, and pose. This study examined hand pose recognition by means of contour-based method proposed by Keogh using depth images obtained from a single depth sensor.

We conducted experiments on recognizing 23 hand poses from 400 Japanese sign language words. Under the condition of close-tester, the recognition rate was approximately 90 % for close-dataset, 70 % for open-dataset. In addition, we conducted an experiment to recognize the hand poses of an unknown person by means of discriminators learned from hand poses of other people. The recognition rate dropped considerably because diversities in bone structure of each person's hand generated loose discriminators that are unable to consider the details of contours. We also evaluated the computation time. Regarding close-tester and open-dataset, 88 ms (corresponding to 11 fps) was required to recognize a hand image with 70 % accuracy.

When we recognize the hand poses of an unknown person, discriminators generated from people who have similar bone structure should be used. Future research in this area requires that hand images of various people be collected and applied for the purpose of recognizing unknown persons.

6. Acknowledgement

This research was supported in part by Japan Society for the Promotion of Science KAKENHI (No. 25350666), and Toukai Foundation for Technology.

7. References

[1] Rung-Huei Liang and Ming Ouhyoung, “A Real-time Continuous Gesture Recognition System for Sign Language,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, Apr 1998, pp. 558–567.

[2] Arata Sato and Koichi Shinoda, “Large Vocabrary Sign Language Recognition Based on Cheremes,” in *IEICE Technical Report PRMU2011-222, SP2011-137*, 2012, pp. 155–160.

[3] Eamonn Keogh, Li Wei, Xiaopeng Xi, Sang-Hee Lee and Michail Vlachos, “LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures,” in *32nd International Conference on Very Large Data Bases (VLDB2006)*, 2006, pp. 882–893.

[4] Kinect for Windows, http://kinectforwindows.org.

[5] Tsutomu Kimura, Daisuke Hara, Kazuyuki Kanda and Kazunari Morimoto, “Expansion of the System of JSL-Japanese Electronic Dictionary: An Evaluation for the Compound Research System,” in *Proceedings of the 2nd International Conference on Human Centered Design*, ser. HCD'11, 2011, pp. 407–416.

[6] Hui Liang, Junsong Yuan and Daniel Thalmann, “Parsing the Hand in Depth Images,” *Multimedia, IEEE Transactions on*, vol. 16, no. 5, pp. 1241–1253, Aug 2014.

[7] Danhang Tang, Tsz-Ho Yu and Tae-Kyun Kim, “Real-Time Articulated Hand Pose Estimation Using Semi-supervised Transductive Regression Forests,” in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ser. ICCV '13, 2013, pp. 3224–3231.

[8] Ming-Kuei Hu, “Visual Pattern Recognition by Moment Invariants,” *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179–187, February 1962.

Synthesizing and Evaluating Animations of American Sign Language Verbs Modeled from Motion-Capture Data

Matt Huenerfauth [1], *Pengfei Lu* [2], *Hernisa Kacorri* [2]

[1] Rochester Institute of Technology, Golisano College of Computing and Information Sciences
[2] The Graduate Center, CUNY, Doctoral Program in Computer Science
matt.huenerfauth@rit.edu, pengfeilv@gmail.com, hkacorri@gradcenter.cuny.edu

Abstract

Animations of American Sign Language (ASL) can make information accessible for many signers with lower levels of English literacy. Automatically synthesizing such animations is challenging because the movements of ASL signs often depend on the context in which they appear, e.g., many ASL verb movements depend on locations in the signing space the signer has associated with the verb's subject and object. This paper presents several techniques for automatically synthesizing novel instances of ASL verbs whose motion-path and hand-orientation must accurately reflect the subject and object locations in 3D space, including enhancements to to prior state-of-the-art models. Using these models, animation generation software could produce an infinite variety of indicating verb instances. Using a corpus of motion-capture recordings of multiple performances of eight ASL indicating verbs, we modeled the signer's hand locations and orientations during each verb, dependent upon the location in the signing space where the subject and object were positioned. In a user study, ASL signers watched animations that included verbs synthesized from these models, and we found that they had similar quality to those produced by a human animator.

Index Terms: American Sign Language, accessibility for people who are deaf, animation, natural language generation

1. Introduction

This paper describes technologies for automating the creation of animations of American Sign Language (ASL), which is a natural language that consists of movements of the hands, body, head, and face. ASL is the primary means of communication for over 500,000 people in the United States [18]. ASL is a natural language, and the grammar, word-order, and vocabulary of the language is distinct from English. For various language-exposure and educational reasons, many deaf adults have lower literacy levels. In fact, standardized testing suggests that the majority of deaf high school graduates in the U.S. (typically students age 18) have a fourth-grade reading level or below (typically students age 10) [22].

Given these literacy trends, when English text is presented online, the text may sometimes be too difficult for many of these users. Conveying information content through videos or animations of ASL could make information more accessible. As discussed in [14], because a human signer must be re-filmed, videos are ill-suited to contexts where the information: is often modified, might require later corrections, is generated automatically in response to a query, or is produced by automatic translation technologies. Animations of sign language that are produced automatically from an easy-to-update script can overcome these limitations and make it easier to incorporate ASL content on websites or other media. A challenge is that ASL signs must be "customized" so that they are performed in a specific manner that matches how the signer has set up locations around their body to represent entities under discussion. This paper focuses on a ubiquitous class of ASL verbs, called "indicating verbs," and it describes research on technologies to automatically produce understandable animations of these verbs for use in ASL animations, with an ultimate goal of increasing information accessibility for users who are deaf.

1.1. Spatial Reference Points, Indicating Verbs

In ASL, entities under discussion (concepts, people, etc.) can be associated with locations in the 3D space around the signer's body [10, 16, 17]. For example, the signer may point to a location in space immediately after an entity is mentioned for the first time; when signers want to refer to the entity again, they do not repeat its name. Instead, they point to the location in space. Some linguists (e.g., [16, 17]) have proposed that ASL signers can be thought of as using a semi-circular arc (at chest height around their torso) as the range of possible locations where entities may be established. The location of the spatial reference point for an entity could be represented as an angle on this arc. Individual signs may vary in how they are performed in a particular sentence, based on a variety of linguistic factors. For instance, temporal aspect, manner, or spatial depiction can be conveyed through modifications to the performance of an ASL verb [4, 5]. However, the focus of this paper is a class of ASL verbs referred to as "indicating verbs" by [10] (also known as "inflecting verbs" [19] or "agreeing verbs" [1].) The movement path and hand orientation of these verbs is affected by the spatial reference points for the verb's subject and/or object [10, 19].

When a signer is asked to perform an indicating verb in isolation or when such a verb is listed in a dictionary, the prototypical verb performance that is seen is typically referred to as a "citation form" or "uninflected form," which has not been modified to indicate locations in the signing space for its subject or object. When an indicating verb is performed in a sentence, the signer will modify the hand locations and orientations used to perform the verb, often tracing a unique motion-path through the signing space, which indicates the locations of the spatial reference points for the verb's subject and/or object. In fact, in ASL sentences that include an indicating verb, the subject or object is often not overtly expressed. That is, the signer does not need to point to the spatial reference locations for the subject or object as long as the verb's motion-path and orientation reveals the identity of its subject and object. If a signer does choose to explicitly mention the subject and object of the verb, then it is legal for

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 22–28,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

the signer to simply use the uninflected form of the verb, but the resulting sentences may appear less fluent. Signers who view ASL animations find those that include citation forms of indicating verbs more difficult to understand (as compared to versions of animations in which indicating verbs indicate the locations of the subject and object) [7].

Generally, the motion path of indicating verbs moves away from the subject and toward the object, but the verb performance is actually a complex interaction of: (a) the verb's citation-form motion-path and hand orientation, (b) the location of the subject's spatial reference point, and (c) the location of the object's spatial reference point. ASL verbs can be partitioned into multiple classes, based on whether their motion is modified based on: (1) subject only, (2) object only, (3) both, or (4) neither [10, 19]. Figure 1 shows the verb EMAIL, which is a verb of type (3).

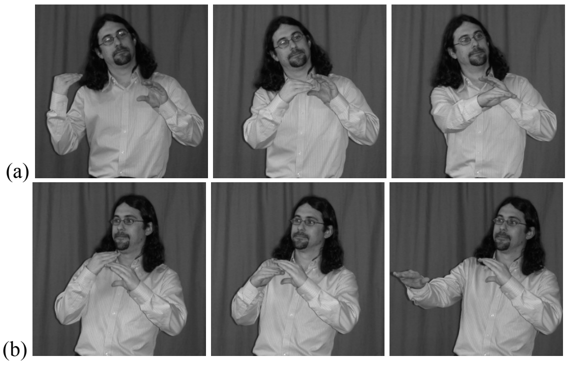


Figure 1: Verb EMAIL with: (a) subject on the left and object on right or (b) with opposite arrangement.

This paper describes our research on automating the creation of animations of ASL indicating verbs. Section 2 briefly summarizes some prior work on modeling ASL indicating verbs. Section 3 describes new techniques for automatically synthesizing animations of ASL verb signs. Section 4 presents an experiment with 18 native ASL signers who evaluate animations resulting from our modeling techniques. Finally, section 5 presents conclusions and avenues for future work.

2. Prior Work on ASL Verbs

Researchers have investigated methods to speed the creation of sign language animations. *Scripting* systems, e.g., [23], allow a human who is knowledgeable of ASL to assemble sentences by drawing upon pre-built words in a dictionary to create a timeline for a performance. A common limitation is that the user may not find the exact sign (or version of a sign) that is needed for a particular sentence, e.g., most systems include only the citation form of verb signs because it is not practical to include hundreds of versions of each verb for various possible arrangements of the verb's subject and object in the signing space. As discussed in [6], other researchers have focused on building *generation* systems, which further automate the production of animation, e.g. research on machine translation of written text into sign language animation. In order for the machine translation output to include indicating verbs, some method is needed for automatically predicting how the motion-path and orientation of a verb would be affected by the locations of the verb's subject and object in space. Sections 2.1 and 3 describe research on automatically synthesizing novel performances of ASL verb signs for any desired combination of subject and object arrangement in the signing space: Such software would be useful in both scripting and generation systems, thereby making it easier to add indicating verbs to animations.

Marshall and Safar [15] designed an animation generator that could associate entities with up to six locations in the signing space and produce British Sign Language verbs whose subject/object were positioned at these locations. However, the verbs involved simple motion paths, and the system did not allow for the arrangement of subjects and objects at arbitrary locations in the signing space (a small number were enabled).

Some researchers have studied videos of performances of ASL verbs to design algorithms for specifying how the arms should move for specific arrangements of subject and object [21]. While the results were promising, a human animation programmer was needed to design the necessary algorithms. By contrast, our research is based on the idea that the only input should be a set of examples of how an ASL verb is performed for various given arrangements of subject and object, with the software automatically learning a model of how a signer's hands should move, given where the subject and object is located in space.

Other researchers have collected motion-capture recordings of signing and used this data to synthesize novel verb signs: Duarte and Gibet [2] collected French Sign Language data via motion capture, and they reassembled elements of the recordings to synthesize novel animations. They used several "channels" to represent their recorded signs, e.g., channels of eye, head, spine, and arms, and they mixed information from the channels of different recordings to produce new animations. For a small number of verb signs, these researchers played the recording of the verb in reverse (from the original recording) to produce a version of the verb with the subject and object in opposite locations. For example, they recorded several indicating verbs with a few combinations of subject/object, e.g., "I invite-you" and "you-invite-I." However, they did not try to build a model of how to synthesize novel inflections for verbs for any arrangement of subject or object in the signing space (the focus of this paper).

2.1. Earlier Work on ASL Indicating Verb Modeling

In earlier work, researchers have designed data-driven methods for synthesizing animations of ASL indicating verbs, for any desired arrangement of the subject and object on an arc around the signer. However, there were significant limitations in that prior work, which we address with some novel modeling approaches described and evaluated in this paper.

As described in [11], verb performances were collected from human ASL signers to create a training data set for animation modeling research. The data included the location (x, y, z) and orientation (*roll*, *pitch*, *yaw*) for the hands, torso, and head of the signer. The native ASL signer performed ASL verbs signs, for given arrangements of the subject and object in the signing space. Targets were positioned around the laboratory at precise angles, relative to where the signer was seated, corresponding to positions on an arc around the signer's body. The signer was asked to perform ASL verbs, e.g., EMAIL, with one target as subject and another as object. In this way, 42 examples of verb forms were recorded for each verb, for various combinations of subject and object locations. Because the verbs considered contained relatively straight motion paths for the hands, they were modeling using two keyframes (one at the beginning of each hand's motion path and one at the end).

Thus, the location (x, y, z) and orientation (*roll*, *pitch*, *yaw*) for the hands were extracted at each keyframe. (For signs with more complex paths, additional keyframes might be required.) This data was used to learn a model to predict the motion-path of a signer's hands for that verb, for novel arrangements of the subject and object on the arc around the signer. In prior work [8, 12, 13], two major types of modeling approaches were created for ASL indicating verbs:

Point-Based Modeling: This model [8, 12] predicted a starting location and the ending location of the hands for the verb, as distinct points in the 3D signing space; the virtual human animation software interpolated between these location points. Based on the position on the arc around the signer where the subject and object of the verb were located, the coefficients of six polynomial models were "fit" from training data for each for each hand (x_{start}, y_{start}, z_{start}, x_{end}, y_{end}, z_{end}), and, at run time, the models were used to estimate these values to synthesize a particular verb instance that was needed for an animation [8].

Vector-Based Modeling: The "point" model was not ideal: When different human signers perform a verb (e.g., EMAIL with subject at arc position on the left and object at arc position on the right), not all of the humans select exactly the same 3D point for their hands to start and stop. What is common across the performances is the *direction* that the hands move through space. Thus, in [13], a new modeling approach was proposed, called "vector" based modeling. Each verb was modeled as a tuple of values: the difference between the x-, the y-, and the z-axis values for the starting and ending location of the hand. Using this model, researchers followed a similar polynomial fitting technique summarized in [8], except that the model used fewer parameters. The "vector" model used only three values per hand ($delta_x$, $delta_y$, $delta_z$), instead of six per hand in the prior "point" model, which represent start and end location of the hand as (x_{start}, y_{start}, z_{start}, x_{end}, y_{end}, z_{end}). Of course, knowing the direction that the hands should move is insufficient: to create an animation, the starting and ending locations for the hands must be selected. At run time, a Gaussian mixture model of hand location likelihood (that had been trained for each ASL verb) was used to select the starting position for each hand (to identify a path that travels through a maximally-likely region of space) to synthesize a particular verb instance for an animation [13].

3. Novel Modeling Approaches

Limitations of prior work ASL verb modeling included:

- While hand orientation (*roll, pitch, yaw*) was modeled using artificially produced testing-data from a human animator in [8], researchers never attempted to model the orientation (*roll, pitch, yaw*) of the hands, based on a training set of motion-capture data *from humans*. Since hand orientation must be selected when producing an ASL animation, this was a major limitation of prior work.
- The "vector" model in prior work treated the left and right hands of the signer as completely independent motion vectors that needed to be selected. Section 3.1 will discuss how this led to low quality animation results for some verbs, and it will address this limitation.

Researchers had never before conducted a user-based evaluation (with native ASL signers viewing animations and answering questions) to compare the point-based and vector-based modeling approaches for synthesizing verbs. This paper presents the first user-based comparison of the quality and understandability of verbs synthesized by those two verb models, trained on motion-capture data from human signers. In addition to the conduct of the user-based study (section 4), another novel aspect of this paper is that we have enhanced and modified the Vector-Based Model, that was first described by [13], in several new ways, as described below.

3.1. Relative Hand Location Modeling

Some ASL verbs involve a movement in which the two hands come into close proximity or interact in a specific spatial orientation. For example, when performing the verb EMAIL as seen in Figure 1, the right hand must pass through the "C" handshape of the left hand. This close-proximity articulation of the two hands is essential for this verb's understandability. As another example, the ASL verb COPY requires the signer's two hands to come into close proximity at the beginning of the performance, as seen in Figure 2 and Figure 3.

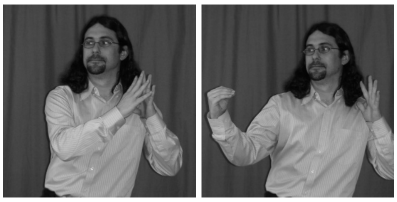


Figure 2: Inflected version of ASL verb COPY with the subject on right and the object on left.

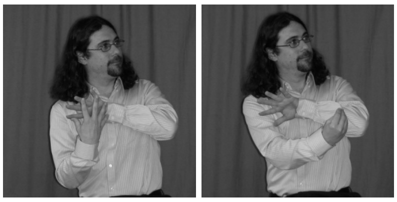


Figure 3: Inflected version of ASL verb COPY with subject on left and object on right.

There is a limitation in the original vector-based model, defined in [13]: That model did not explicitly represent the *relative* location between the left and right hands. It represented the direction of each hand's movement, with the starting location of each hand selected *independently,* based on the Gaussian model of hand location likelihood for that ASL verb. Applying such a technique to several examples of verbs such as EMAIL and COPY with specific hand proximity requirements, it was apparent that independently modeling the direction of both the left and right hands led to animations in which the relative positions of the two hands were not correctly preserved during the performance of the verb. For instance, for a verb like EMAIL, the right hand did not always move precisely through the opening produced by the left hand.

Therefore, we re-implemented and modified that original vector-based model, as follows: we model the left hand position *relative to* the right hand's position at each keyframe of the verb. At run time, we used our model to predict a hand movement direction vector for the right hand only. When we needed to synthesize a specific verb instance, we first selected a right hand starting location based on the Gaussian model. Then, we used our model of left hand relative-location to select a left hand location for each key-frame, *relative* to the right hand. Our new vector-based model, for verbs with two keyframes, would model nine values ($delta_x$, $delta_y$, $delta_z$) for the right hand and (*relative*$_x$, *relative*$_y$, *relative*$_z$) for the left hand for each keyframe of the verb. In the prior "point" model, for a two-keyframe verb, there would be a total of twelve values modeled, the start and end location of both hands as (x_{right}, y_{right}, z_{right}, x_{left}, y_{left}, z_{left}). Given this new vector-based model (with the left-hand locations represented as relative to the right hand locations), we trained our enhanced vector-based models on the motion-capture data of ASL verbs that had been recorded by prior ASL animation researchers [14].

3.2. Modeling Hand Orientation

In prior work, researchers had not modeled *orientation* of the hands for ASL verbs using motion-capture data collected from human signers [13]. In this section, we present a novel method for modeling hand orientation (and an evaluation in section 4). Because there are various popular methods of representing the orientation of 3D objects (e.g., Euler angles, axis-angle, or 3x3 rotation matrices), we had to select an approach that was well-suited to representing hand orientation for modeling ASL verbs. Almost all orientation representations are actually representations of the 3D *rotation* of an object from a starting orientation; they all assume that a 3D object enters the universe with some initial orientation. They differ as follows:

- *Euler angles* represent a sequence of three rotations about the local axes of an object in 3D space. For instance, a first rotation about the object's z-axis by an angle α, a second rotation about its x-axis (which might have been affected by the first rotation) by an angle β, and another rotation about the object's z-axis, by an angle γ [3].
- The *axis-angle* representation is a rotation representation that consists of a unit vector <*x*, *y*, *z*> indicating an axis of rotation in a three-dimensional space and an angle *theta* indicating the magnitude of the rotation [3].
- A *rotation matrix* is another way to represent orientations of 3D objects; in this case, a 3x3 matrix can be used to represent a rotation. To rotate a point in three-dimensional space (represented as column vectors), you can multiply it by the 3x3 rotation matrix [3].

Since there are methods for converting between various orientation representations, we were free to select whichever representation for our modeling of hand orientation of ASL verbs. We wanted to select an approach with desirable mathematical properties. Specifically, we prefer methods of modeling orientation that avoid gimbal lock (described below) and were well suited to interpolation (meaning that when you numerically average the numbers that represent the orientation, the resulting 3D orientation of the object looks realistic). Techniques for computing representative orientations from measured 3D data have been described by several researchers, e.g., [5, 24], and the relative tradeoffs of many of these techniques have also been investigated, e.g., [25]. Some relevant considerations are summarized below:

- If we had used Euler angles, we may have encountered problems due to gimbal lock, a phenomena in which the first Euler rotation causes the axes of the system to align in such a way that a degree of freedom is lost [3].
- If we had used axis-angle representations, we may have encountered problems because axis-angle representations are not a unique representation of orientation (meaning that there are multiple possible ways to represent the same resulting final orientation of an object). Thus, there is no guarantee that simple interpolation of the numbers of the orientation representation will result in a realistic-looking 3D orientation for the final object (because the resulting orientation produced through interpolation may not be on the shortest path on the great arc between the two original orientations).
- If we had used 3x3 rotation matrices to represent orientation for modeling, this would have made our modeling more complex because this representation uses a large number of parameters (specifically, nine) to represent orientation.

For these reasons, we selected a less common method of representing orientations: Simultaneous Orthogonal Rotation Angles (SORA). SORA represents a rotation as a vector of three values (φ_x, φ_y, φ_z) that represent three *simultaneous* rotations around the coordinate system axes. (Euler angles represent *sequential* rotations.) SORA has been used in the areas of real-time angular velocities estimation [20]. The simplicity of SORA makes it possible for our orientation modeled in a single step, and avoids several of the problems with other approaches, outlined above. There are also standard ways to convert between SORA and other orientation representations [11, 20]. While [25] identify some limitations of SORA (similar to discontinuities encountered with axis-angle), we have found SORA to be an effective modeling approach for ASL verb orientation (as shown in Section 4.)

We performed our modeling as follows: First, we converted the motion-capture data into SORA format. Then, we trained the orientation models for all eight verbs (TELL, SCOLD, GIVE, MEET, ASK, EMAIL, SEND, and COPY). Since the rotation component for each axis can be isolated when using SORA, we consider the axes independently when we fit 3rd order polynomials to predict each component of SORA. Figure 4 outlines the procedure. At run-time, given some *s* and *o* values (i.e., subject and object location on the arc around the signer), we independently predict each of the values of φ_x, φ_y, and φ_z. After modeling each SORA value, we converted this back to axis-angle to synthesize a verb animation.

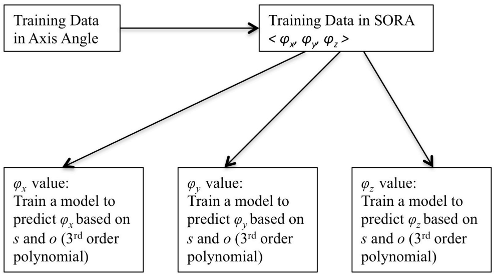

Training Data in Axis Angle
Training Data in SORA < φ_x, φ_y, φ_z >
φ_x value: Train a model to predict φ_x based on *s* and *o* (3rd order polynomial)
φ_y value: Train a model to predict φ_y based on *s* and *o* (3rd order polynomial)
φ_z value: Train a model to predict φ_z based on *s* and *o* (3rd order polynomial)

Figure 4: Training verb orientation data using SORA.

4. USER-BASED EVALUATION STUDY

A user study was conducted to evaluate animations synthesized by our point-based model and by our vector-based model, trained on the recorded data of the eight ASL verbs. The overall methodology of this study, including the recruiting practices, format of comprehension questions, and other details follows the general approach used in prior ASL evaluation research, e.g., [8]. Of the 24 participants, 13 had used ASL since infancy, 6 participants had learned ASL before age 8, and 2 participants began using ASL at a school with primary instruction in ASL since age 10. The remaining 3 participants identified as deaf, attended schools and university with instruction in ASL, and had spouses or partners with whom they used ASL on a daily basis. There were 17 men and 7 women of ages 24-58 (median age 33).

The experiment consisted of two phases: In phase 1 of the study, we used a set of 12 ASL stories and comprehension questions that we designed and produced as stimuli. The stories and questions were adapted from those used in [8] for use in this current study; the stories were edited so that they included the eight ASL verbs listed in Table 1. The animations consisted of a single onscreen virtual human character, who tells a story about 3-4 characters, who are associated with different arc-positions in the signing space surrounding the virtual signer. The 12 stories and their questions were designed so that the questions related to information conveyed by a specific verb in the story. The comprehension questions were difficult to answer because of the stories' complexity, because participants saw the story before seeing the questions, and because they could only view the story one time. Each story was produced in four different versions, based on the form of the verb used in the animation:

- PointModel: inflected verb using our point-based model
- VectorModel: inflected verb using vector-based model
- Animator: inflected verb produced by a human animator
- Uninflected: uninflected citation-form of the verb

It is important to note that all of the animations presented were ***grammatical***, including the Uninflected stimuli. As described in section 1.1, verbs in ASL do not require spatial inflection during sentences, so long as the identity of the subject and object is otherwise indicated in the sentence. The animations presented in this study included in this information in the form of noun phrases or pointing pronouns in each sentence, identifying the subject and object. So, there were no non-grammatical sentences shown to participants in the study.

Section 3.2 mentions how the orientation model of the vector-based model is identical to the orientation model of the point-based model, so, the hand orientations in these two types of animation are identical – only the locations of the hands differ.

In this within-subjects study design:

- No participant saw the same story twice.
- The order of presentation of each story was randomized.
- Each participant saw 3 animations of each version.

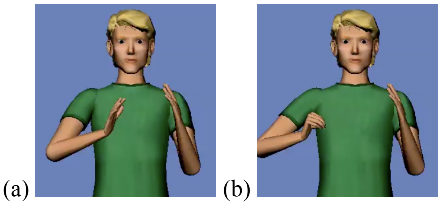


Figure 5: Example of ASL verb COPY produced by the vector model, as it appears in the study

Figure 5 shows example images for the verb COPY, produced by the vector model, as it appeared in a story during the study. In this example, the animated signer described a story in which several students (set up at locations in the signing space) were working on homework, and one student copied another student's homework. One of the comprehension questions for this story asked which of the students copied the homework. Animation examples from this study may be accessed here: http://latlab.ist.rit.edu/slpat2015/

Table 1: Verbs collected in the training data set and which appear in the stimuli in study in section 4.

Verb	***Indicates***	***1- or 2-handed***	***Description of Movement***
ASK	Subject and Object	1	'ask a question': a bending index finger moves from Subj ('asker') to Obj ('askee')
GIVE	Subject and Object	2	'give to someone': hands move as a pair from the Subj ('giver') to Obj ('recipient')
MEET	Subject and Object	2	'two people meet': hands move from Subj and Obj toward each other, and meet somewhere in-between.
SCOLD	Object only	1	'scold/reprimand': extended index finger wags at the Obj ('person being scolded')
TELL	Object only	1	'tell someone': index finger moves from signer's mouth to Obj ('person being told')
COPY	Subject and Object	2	'copy from someone': right flat hand against left flat hand near Obj ('someone') moves toward Subj ('copier').
EMAIL	Subject and Object	2	'email to someone': right hand (bent-flat) passed through the cavity of the left hand (C shape) from Subj to Obj.
SEND	Subject and Object	2	'send to someone': a "B" hand with fingertips' quickly slide over the back of other hand, moving from Subj to Obj.

After watching each story once, participants answered 4 multiple-choice comprehension questions that focused on information conveyed by the indicating verbs. This study followed the methodological details of prior ASL animation research studies, as described in [8, 9, 11]. Figure 6 shows the comprehension question accuracy scores. A Kruskal-Wallis test (alpha=0.05) was run to check for significant differences between comprehension scores for each version of the animations. Only one pair of values had a significant difference (marked with a star in the Figure).

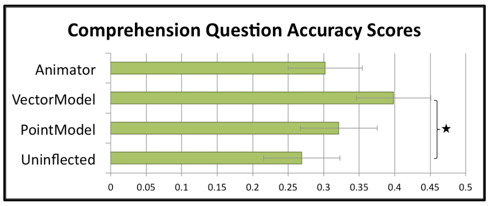


Figure 6: Comprehension question scores in phase 1.

In phase 2, participants viewed four animations of the same sentence side-by-side; e.g., "John point_to_arc_position_0.9 ASK Mary point_to_arc_position_-0.6." (Arc position 0.9 is on the signer's far right side, and arc position -0.6 is on the signer's left side.) The only difference between the four versions that were displayed on the screen was whether the *verb* in the sentence was: (a) synthesized from our point-based model, (b) synthesized from our vector-based model, (c) created by a human animator, or (d) an uninflected version of the verb. Participants could re-play the animations multiple times, and a variety of arc-positions were used in the animations (the four versions shown at one time all used the same arc-positions). Participants answered 1-to-10 Likert-scale questions about the quality of the verb in each of the 3 versions of the sentence. Figure 7 shows the results. To check for significant differences between Likert-scale scores for each version, a Kruskal-Wallis test (alpha=0.05) was performed; significant pairwise differences are marked with a star.

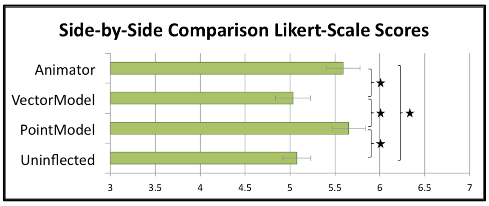


Figure 7: Subjective Likert-scale scores in phase 2.

4.1. Discussion of Results

For the comprehension question scores collected in phase 1 of the study, the vector-based model had significantly higher scores than the stories with the uninflected version of the verbs. This is a positive result because it indicates that the vector-based modeling approach led to more understandable stories. Prior work [9] has shown that comprehension-question based evaluation of animations is necessary to accurately measure the understandability of ASL animations.

For the subjective scores of animation quality collected during the side-by-side comparisons in phase 2 of the study, the animations containing verbs produced by the human animator received significantly higher scores than the uninflected animations. This was an expected result: the Animator animations were hand-crafted by a native ASL signer with proper ASL verb inflection movements, whereas the Uninflected animations were considered our lower baseline.

Similar to the Animator animations, our PointModel animations received higher subjective evaluation scores than the Uninflected animations. Verbs produced using this modeling technique received higher scores from native ASL signers. Uninflected verb animations are still used in many sign language animation systems; so, this indicates that our modeling technique is superior to that lower baseline. Because the Animator version of the verbs was considered our upper baseline for this study (since it reflects the careful creation of an inflected verb form during a time-consuming process), it was a positive result that the PointModel achieves this high score.

It is also notable that the PointModel received statistically higher subjective scores than the VectorModel, and the VectorModel did not receive statistically higher scores than the Uninflected animations. This result may indicate that there were problems with some animations produced using the VectorModel in this study. Figure 8 shows per-verb results from phase 2. It is important to note that none of the differences in Figure 8 were statistically significant; however, looking at this figure, we *speculate* that the VectorModel may have performed poorly for TELL and SCOLD. Among the verbs in this study, these two verbs are special, in that they inflect for object position only. (Their movement path is not modified based on where the subject of the verb is positioned on an arc around the signer.) Further, when human signers perform these verbs, their motion path is oriented away from the signer's chin (in the case of TELL) or heart (in the case of SCOLD). Since the VectorModel does not explicitly model the starting location of a verb (the location is selected based on a search through the Gaussian mixture model representing hand location probability), the VectorModel may lead to verb animations in which the starting location is somewhat inaccurate. For some ASL verbs, this may not have a significant impact on the perceived quality of the verb, if the overall direction of the verb movement is correct. However, for TELL and SCOLD, it may be the case that the beginning location of these verbs is very important for the correct production of the sign. For this reason, the vector model may not be appropriate for verbs of this type. Investigating the suitability of the vector model for different classes of ASL verbs, that have particular constraints on their starting locations, is an open area of future research.

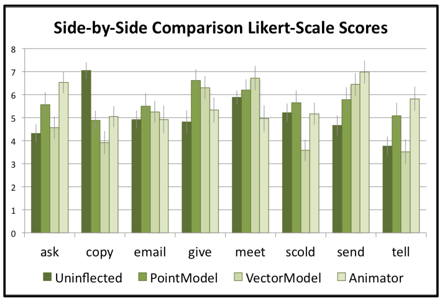


Figure 8: Per-verb results from phase 2 of the study.

5. Conclusions, Future Work

This paper has described our modeling methods and construction of a parameterized lexicon of ASL verb signs, whose motion path depends on the location in the signing space associated with the verb's subject and object. Specifically, we have described enhancements (representing hand orientation and relative location of the hands) to two prior state-of-the-art methods for generating ASL indicating verb animations (i.e., the point-based model and vector-based model of [8, 11, 13]). We have used motion capture data of sign language performances from native signers as a training data set for learning our models. In a user-based evaluation with 24 participants, we evaluated whether these models were able to produce more understandable ASL verb animations.

In future work, we intend to collect a larger set of recordings of ASL indicating verbs, including some with more complex movements of the hands, to evaluate whether the modeling techniques perform well for an even larger variety of signs. We may also explore how subject/object locations affect the signer's handshape during a verb signs: handshape was not affected by subject/object location in our current modeling approaches. We will study how the speed or timing of verb movements varies with the location of subject/object in the signing space. While our current work has focused on verb signs, we believe these modeling techniques may also be applicable to ASL pronouns and other signs whose movements are affected by the arrangement of spatial reference points in the signing space. Further, while this paper focused on ASL, we expect that researchers studying other sign languages internationally may wish to replicate the data-collection and verb-modeling techniques to produce models for signs that are affected by spatial locations.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0746556 and 1506786. Jonathan Lamberton and Miriam Morrow recruited participants and collected response-data during the user-based evaluation study.

7. REFERENCES

[1] Cormier, K. 2002. Grammaticalization of indexic signs: How American Sign Language expresses numerosity. Ph.D. Dissertation, University of Texas at Austin.

[2] Duarte, K., Gibet, S. 2011. Presentation of the SignCom project. In Proc 1st Int'l Workshop on Sign Language Translation and Avatar Technology, Berlin, Germany, 10-11.

[3] Dunn, F., Parberry, I. 2002. 3D Math Primer for Graphics and Game Development. A K Peters/CRC Press. 2nd edition.

[4] Emmorey, K., Bellugi, U., Friederici, A., Horn, P. 1995. Effects of age of acquisition on grammatical sensitivity: Evidence from on-line and off-line tasks. Applied Psycholinguistics, Cambridge University Press. 16(1):1-23

[5] Gramkow, C. 2001. On Averaging Rotations. Journal of Mathematical Imaging and Vision 15: 7–16, 2001. Netherlands: Kluwer Academic Publishers.

[6] Huenerfauth, M., Hanson, V. 2009. Sign language in the interface: access for deaf signers. In C. Stephanidis (ed.), Universal Access Handbook. NJ: Erlbaum. 38.1-38.18.

[7] Huenerfauth, M., Lu, P. 2012. Effect of spatial reference and verb inflection on the usability of sign language. Universal Access in the Information Society 11(2):169-184.

[8] Huenerfauth, M., Lu, P., 2010. Modeling and synthesizing spatially inflected verbs for American sign language animations. In Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (ASSETS '10). ACM, New York, NY, USA, 99-106.

[9] Huenerfauth, M., Zhao, L., Gu, E., Allbeck, J. 2007. Evaluating American Sign Language generation through the participation of native ASL signers. In Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility (Assets '07). ACM, New York, 211-218.

[10] Liddell, S. 2003. Grammar, Gesture, and Meaning in American Sign Language. UK: Cambridge U. Press.

[11] Lu, P. 2013. Data-driven synthesis of animations of spatially inflected American Sign Language verbs using human data. Ph.D. dissertation, City University of New York, NY, USA.

[12] Lu, P., Huenerfauth, M. 2011. Synthesizing American Sign Language spatially inflected verbs from motion-capture data. In Proceedings of the 2nd International Workshop on Sign Language Translation and Avatar Technology (SLTAT), in conjunction with ASSETS 2011, Dundee, Scotland.

[13] Lu, P., Huenerfauth, M. 2012. Learning a vector-based model of American Sign Language inflecting verbs from motion-capture data. In Proceedings of the 3rd Workshop on Speech and Language Processing for Assistive Technologies (SLPAT '12). ACL, Stroudsburg, PA, USA, 66-74.

[14] Lu, P., Huenerfauth, M. 2014. Collecting and evaluating the CUNY ASL corpus for research on American Sign Language animation. Computer Speech & Language 28(3):812–831.

[15] Marshall, I., Safar. E. 2005. Grammar development for sign language avatar-based synthesis. In Proc. UAHCI'05.

[16] McBurney, S.L. 2002. Pronominal reference in signed and spoken language. In R.P. Meier, K. Cormier, D. Quinto-Pozos (eds.) Modality and Structure in Signed and Spoken Languages. UK: Cambridge U. Press, 329-369.

[17] Meier, R. 1990. Person deixis in American Sign Language. In S. Fischer, P. Siple (eds.) Theoretical issues in sign language research. Chicago: U. Chicago Press, 175-190.

[18] Mitchell, R., Young, T., Bachleda, B., & Karchmer, M. 2006. How many people use ASL in the United States? Why estimates need updating. Sign Lang Studies, 6(3):306-335.

[19] Padden, C. 1988. Interaction of morphology & syntax in American Sign Language. New York: Garland Press.

[20] Stančin, S., Tomažič, S. 2011. Angle estimation of Simultaneous Orthogonal Rotations from 3D gyroscope measurements. Sensors 2011, 11, 8536-8549.

[21] Toro, J. 2005. Automatic verb agreement in computer synthesized depictions of American Sign Language. Ph.D. dissertation, DePaul University, Chicago, IL.

[22] Traxler, C. 2000. The Stanford achievement test, 9th edition: national norming and performance standards for deaf & hard-of-hearing students. J Deaf Stud & Deaf Educ 5(4):337-348.

[23] VCom3D. 2014. Homepage. http://www.vcom3d.com/

[24] Pennec, X., FIllard, P., Ayache, N. 2006. A Riemannian Framework for Tensor Computing. International Journal of Computer Vision 66(1):41–66, January 2006, Springer.

[25] Allgeuer, P., Behnke, S, 2014. Fused Angels for Body Orientation Representation. In proceedings of the 9th Workshop on Humanoid Soccer Robots, IEEE-RAS International Conference on Humanoid Robots (Humanoids), Madrid, Spain, 2014.

Evaluating a Dynamic Time Warping Based Scoring Algorithm for Facial Expressions in ASL Animations

Hernisa Kacorri[1], *Matt Huenerfauth*[2]

[1]The Graduate Center, CUNY, Doctoral Program in Computer Science, USA
[2]Rochester Institute of Technology, Golisano College of Computing and Information Sciences, USA

hkacorri@gradcenter.cuny.edu, matt.huenerfauth@rit.edu

Abstract

Advancing the automatic synthesis of linguistically accurate and natural-looking American Sign Language (ASL) animations from an easy-to-update script would increase information accessibility for many people who are deaf by facilitating more ASL content to websites and media. We are investigating the production of ASL grammatical facial expressions and head movements coordinated with the manual signs that are crucial for the interpretation of signed sentences. It would be useful for researchers to have an automatic scoring algorithm that could be used to rate the similarity of two animation sequences of ASL facial movements (or an animation sequence and a motion-capture recording of a human signer). We present a novel, sign-language specific similarity scoring algorithm, based on Dynamic Time Warping (DTW), for facial expression performances and the results of a user-study in which the predictions of this algorithm were compared to the judgments of ASL signers. We found that our algorithm had significant correlations with participants' comprehension scores for the animations and the degree to which they reported noticing specific facial expressions.

Index Terms: American Sign Language, accessibility for people who are deaf, animation, natural language generation

1. Introduction

Access to understandable information on websites and other media is necessary for full participation in society. Yet, the vast majority of information content online is in the form of written language text, and there are many users who have difficulty reading this material. For many people who are deaf and hard-of-hearing, there are educational factors that may lead to lower levels of written language literacy. In the U.S., standardized testing has revealed that a majority of deaf high school graduates (students who are age 18 and older) have a fourth-grade English reading level or below [27]. (U.S. students in the fourth grade of school are typically age 10.) While they may have difficulty with written English, many of these users have sophisticated fluency in another language: American Sign Language (ASL).

More than 500,000 people in the U.S. use ASL as a primary means of communication [20]. However, fluency in ASL does not entail fluency in written English since the two are distinct natural languages: with their own word order, linguistic structure, and vocabulary. Thus, information content can be easier to understand for many deaf users if it is presented in ASL. A spontaneous approach to presenting ASL online would be to upload videos of human signers on website and other media, but this is not ideal: re-filming a human performing ASL for frequently updated information is often prohibitively expensive, and the real-time generation of content from a query is not possible. Software is needed that given an easy-to-update script as input can automatically synthesize ASL signing performed by a virtual human character. This software must internally coordinate the movements of the virtual human character such that the animated ASL message is linguistically accurate, understandable, and acceptable among users. The creation of such software is the focus our research.

An ASL utterance consists of the movement of the hands, arms, torso, head, eye-gaze, and facial expressions. In fact, facial expressions are essential to the understandability and meaning of ASL sentences (see section 2). Our research focuses on the automatic synthesis of facial expression movements for an ASL-signing virtual human character such that the resulting animations are judged to be clear and understandable by deaf users. In addition to our ongoing research in this area, other groups have studied issues related to the synthesis of facial expressions for sign language animation, whose methods and contributions we compare and survey in [14]. For researchers like ourselves, who are interested in designing software that generates linguistically-accurate ASL facial expressions performed by virtual human characters, the most comprehensive way to evaluate the quality of the software is to conduct user studies. Typically, we generate animations using the facial expression selection software, set up an experiment in which deaf participants view and evaluate the animations, and compare the scores of animations produced using the software (to some baselines or to prior versions of the software). Of course, conducting such studies with users is time-consuming and resource-intensive; so, these studies cannot be conducted on a frequent basis (e. g., weekly) during the development of ASL facial-expression synthesis software. For this reason, it would be useful to have some automatic method for quickly evaluating whether the facial expression produced by the software for some specific ASL sentence is accurate. In this paper, we present an automatic scoring algorithm that can compare two facial expression performances to rate their similarity. In principle, this automatic scoring tool could be used to quickly evaluate whether the output of facial expression synthesis software is producing a result that is similar to ASL utterances recorded from actual human ASL signers. The proposed algorithm could be incorporated into a data-driven facial expression synthesis architecture, an approach which is also favored by other sign language animation researchers, e. g.: [26] that use computer vision to extract facial features and produce facial expressions that occur during specific signs, and [3] that map facial motion-capture data to animation blend-shapes using machine-learning methods.

The face and head position of a virtual human character

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 29–35,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

at any moment in time can be conceptualized as a vector of numbers, specifying joint angles and facial-control parameters at that moment in time. Thus, an animation is a stream of such vectors. While there are a variety of techniques that can be used to measure the similarity between two time-streams of vectors, this paper will specifically explore an approach based on a Dynamic Time Warping (DTW) algorithm. Section 5 describes DTW and discusses how some researchers have begun to use this algorithm to rate the similarity of non-sign-language emotional facial expressions for animated characters [19]; however, no user-study had been performed to verify that such scores actually matched human judgments of similarity – nor has this technique yet been applied to sign-language facial expressions.

This paper presents a novel, sign-language specific scoring algorithm based on DTW, which takes into account the timing of words in the sentence. This paper reflects our first efforts at designing a DTW-based scoring tool, and the goal of this paper is to determine if the technique holds promise – if so, then we intend to investigate further variations of the scoring algorithm, to optimize it for ASL. In order to determine if our scoring tool is useful, we must determine whether the scores it provides actually correlate with the judgments of human ASL signers who evaluate ASL animations in an experiment. This paper presents a user study we conducted in which human ASL signers evaluated animations with facial expressions of different levels of quality (as rated by the automatic scoring tool), and we measure how well our automatic scoring correlates with the human judgments.

The remainder of this paper is organized as follows: Section 2 describes the linguistics of various ASL facial expressions, and section 3 describes how we time-warp a motion-capture recording of a facial expression performance to suit the synthesis of an ASL animation of a sentence with a different time duration. Section 4 describes how the movements of the face of a virtual human character can be parameterized and controlled, and Section 5 defines our new DTW-based automatic scoring algorithm. Section 6 presents our research questions and hypotheses, which were evaluated in a user-study presented in section 7. Finally, section 8 discusses these results and identifies future directions.

2. Syntactic facial expressions

Facial expressions are an essential part of the fluent production of ASL. They can convey emotional information, subtle variations in the meaning of words, and other information, but this paper focuses on a specific use of facial expressions: to convey grammatical information during entire syntactic phrases in an ASL sentence. ASL sentences with identical sequence of signs performed by hands can be interpreted differently based on the accompanying facial expressions. For instance, a declarative sentence (ASL: "ANNA LIKE CHEESECAKE" / English: "Anna likes cheesecake.") can be turned into a Yes-No question (English: "Does Anna like cheesecake?"), with the addition of a Yes-No Question facial expression during the sentence. Similarly, the addition of a Negation facial expression during the verb phrase "LIKE CHEESECAKE" can change the meaning of the sentence to "Anna doesn't like cheesecake." where the signing of the word NOT is optional. For an interrogative question (typically including a "WH" word in English such as where, why, and what), e. g. "ANNA LIKE WHAT", a co-occurring WH-Question facial expression is necessary during the ASL sentence. Instances of these three ASL facial expressions are illustrated in Figure 1.

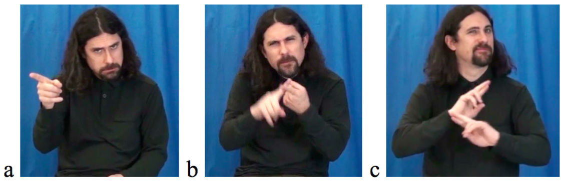


Figure 1: *Examples of ASL linguistic facial expressions: (a) Yes-No Question, (b) WH-Question, (c) Negation.*

While we use the term "facial expressions," these phenomena also include movements of the head, which we model in this paper. ASL linguistics references contain more detail about each, e. g., [22], but a subset of them is described briefly below:

- Yes-No Question: The signer raises his eyebrows while tilting the head forward during a sentence.
- WH-Question: The signer furrows his eyebrows and tilts his head forward during a sentence.
- Negation: The signer shakes his head left and right during the phrase with some eyebrow furrowing.

An ASL linguistic facial expression varies in the way it is performed during a given sentence based on the overall number of signs, the start and end times for a particular word in the sentence (e. g., WHAT and NOT), preceding and succeeding facial expressions, signing speed, and other factors. Thus, simply playing on a virtual character a pre-recorded human performance of a facial expression to a novel, not previously recorded, sentence is insufficient. For this reason, we are investigating how to model and synchronize to manual movements the performance of a facial expression in various contexts.

3. Time-warping facial expressions

In our research on synthesizing ASL animations, we often need to generate a novel animation by assembling a sequence of individual words from a prebuilt animation dictionary; each word may have its own typical duration, which is used to determine a timeline for the full ASL utterance. We seek to add a facial expression performance to such animations, and in section 4, we discuss how facial features extracted from the recording of a human's face could be used to drive the movements of the animated character. Thus, the time-duration of the recording must be "warped" to match the time duration needed in the animation to be synthesized.

Simplistically, the recording could be linearly stretched or squeezed to suit the target time duration, but animation researchers have investigated a variety of techniques for time-warping motion data to new contexts, e. g., [7, 31]. In many approaches, e. g., [7], key milestones during a recorded action are identified in the timestream (e.g., each footfall during a walking action), and these milestone times are used as parameters to determine how to warp the recording (so that the movements of the human for each "footstep" of the walking action are warped into appropriate footstep actions that meet timing requirements for when the virtual human footsteps must occur in the animation).

When synthesizing sign language animations, we have access to information about the underlying timeline of the utterance, which we can use to select useful milestones for time-warping:

- ASL facial expressions occur in relation to the timing of the words during a sentence [22]. Yes-No Question and WH-Question facial expressions typically extend across entire clauses, and Negation, across an entire verb phrase.

- Signers perform anticipatory head movements so that the main action begins with the clause or phrase [22].
- Many phrases with facial expressions begin with or end with a word that has a special relationship to the facial expression being performed (such that there may be additional intensity of the facial expression during this initial/final word).
  - Negated verb phrases may include the word NOT at the beginning of the phrase, where greatest intensity of the Negation facial expression will occur [22].
  - WH-Question clauses typically end with a WH-word, and in some contexts, the facial expression may occur only (or with greatest intensity) during this word [18].
  - Yes-No Question clauses often end with a right-dislocated pronoun [22] or a "QM-wg" (wiggling finger question mark) sign at the end [1].

For an ASL animation that contains a sequence of words, S, when a facial expression occurs, we define four phases of time based on the intervals between five milestones on the timeline:

M1: The end of the word immediately before S

M2: The beginning of the first word in S

M3: For Negation, M3 is the beginning of the second word in S, otherwise, M3 is the beginning of the last word in S

M4: The end of the final word in S

M5: The beginning of the word that immediately follows S

If S begins or ends an utterance, then M1 and M5 are set to a value 500msec away from S. The rationale for these definitions is:

- Phases M1-M2 and M4-M5 represent the onset and offset of the facial expression, before and after S.
- For a Negation phrase, M2-M3 is the duration of the first word, and M3-M4 is the remainder of the phrase. A Negation phrase may begin with the word NOT, when a particularly intense facial expression may occur. Thus, it is useful to distinguish the time of the first word of the phrase. (If S contains only one word, then these phases are merged.)
- For a Yes-No Question or a WH-Question, M3-M4 is the duration of the final word, and M2-M3 is the remainder of the phrase. There is often additional facial expression intensity during the final word of a question; thus, it is useful to distinguish the time of the final word of the question. (If S contains only one word, then these phases are merged.)

Recall that our goal is to modify the timing of a human's facial movement recording to suit the timeline of a target animation we want to synthesize. For any human recording that we plan to use as a source material for facial movements, we ask an ASL signer to identify these five milestones. When we want to modify the timing of a recording, we perform time-warping for each of these four phases independently. Thus, data from phase M2-M3 of the recorded human utterance is time-warped to fit the duration of phase M2-M3 of the target animation that we are synthesizing. In this way, we can increase the likelihood that the appropriate portion of the human's facial performance coincides with the timing of the appropriate signs in the resulting animation.

The top of Figure 2 shows how a recording of the eyebrow height of a human signer during a Yes-No question might appear during an ASL sentence: "SHE LIVE DC SHE" (English: "Does she live in DC?"). The milestones are marked with vertical lines, and the figure shows how data from each phase of the recording can be linearly time-warped to produce a facial expression for an animation with different word durations. (The graph in Figure 2 is an artist's rendering meant to illustrate the warping technique.)

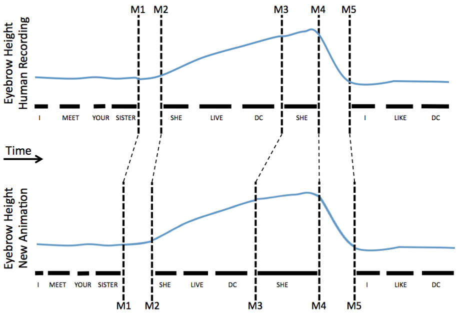


Figure 2: *Phase-based time-warping of a recording of a human's eyebrow movements from a Yes-No Question (above) for an animation with a different timeline (below).*

4. MPEG-4 and ASL animation

In prior work, we constructed a lexicon of ASL signs and a collection of ASL stimuli [9] for use in experiments to evaluate facial expression animation synthesis methods. As part of that project, we recorded videos of a native ASL signer performing the stimuli, and we extracted the facial features and head pose of the human signer in the videos using the Visage Face Tracker (shown in Fig. 3). Visage is an automatic face tracking software [24] that provides a stream of MPEG-4 Facial Action Parameters (FAPs) that represent the facial expression of the human.

The MPEG-4 standard [11] defines a 3D model-based coding for face animation. The facial expression of a human (or an animated character) can be represented by a set of 68 FAPs, representing head motion, eyebrow, nose, mouth, and tongue controls, all of which can be combined for representation of natural facial expressions. For example, "raise_l_i_eyebrow" is one of the FAPs (codename FAP30) in the MPEG4 standard, and it represents the vertical displacement of left inner eyebrow. Larger values for this number would indicate that the eyebrow is raised higher. To specify a changing facial expression over time, a stream of numerical values for all of the FAPs of the face is needed, for each frame of animation.

MPEG-4 FAPs have been used by a variety of non-sign-language animation researchers studying, e. g., expressive embodied agents [21], emotional facial expressions during speech in synthetic talking heads [19], or dynamic emotional expressions [30]. A useful property of MPEG-4 is that the FAP values are normalized to the proportion of the character's face as shown in Fig. 3; thus, a stream of FAP values could be used to drive the animation of virtual humans with different face proportions,and the resulting animation would appear to have similar facial expressions, when played on a difference virtual human.

To support our research on ASL facial expressions (especially the development of automatic scoring tools), it was necessary to implement a virtual human animation platform with face-movement control parameters. We decided to use MPEG-4 facial action parameters [11] , and we enhanced the EMBR platform [5, 6, 16] with MPEG-4-based face controls. We also

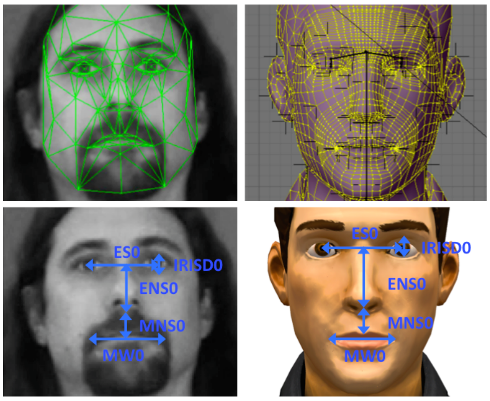


Figure 3: *MPEG-4 facial features and scaling factors on the human signer in Visage (left) and the avatar (right).*

implemented an intermediate component that converts MPEG-4 data to EMBRscript, the script language supported by the EMBR platform. Our script generation component performs the phase-based time-warping approach described in section 3 to align the facial expression with the animated character hand movements. The FAPs that are used to drive our facial expression animations for this paper include the following (additional FAPs may be implemented in future work):

Head orientation (FAP47-FAP49): orientation parameters given in Euler angles defined as pitch, yaw, and roll. In addition to head orientation, the Visage output also includes the head's location in 3D space; we adjust the torso movements of our avatar based on these values.

Vertical displacements of eyebrows (FAP30-FAP35): 6 parameters directly applied the inner, middle, and outer points of the left and right eyebrow to allow for different combinations of raised and lowered eyebrows.

Horizontal displacements of eyebrows (FAP36-FAP37): 2 parameters directly applied in the inner points of the eyebrows that allow for e.g. furrowed eyebrows.

5. The dynamic time warping algorithm

In this paper, we present a novel method for evaluating the quality of synthesized facial expressions for sign-language animations, which is based in the Dynamic Time Warping (DTW) algorithm. DTW arose in the field of speech recognition [25, 28] as a generalization of algorithms for comparing series of values with each other. DTW sums the distance between the individual aligned elements of two time series, which are locally stretched or compressed, to maximize their resemblance. Unlike the Euclidean distance, it can serve as a measure of similarity even for time series of different length. An advantage of DTW over other cross-correlation similarity measures is that it allows for non-linear warping. There are a variety of DTW algorithms, used in several fields, with different global or local constraints (e. g., local slope, endpoints, and windowing), different feature spaces for the time series values, and different local distance metrics between the individual aligned elements (e. g., Euclidean, Manhattan).

DTW has been used as a similarity scoring technique for facial animation, e. g., for the retrieval of facial animation based on a key-pose query [23] and spatio-temporal alignment between face movements recorded from different humans [31]. In prior work, DTW has been also considered as a method for scoring the quality of time series data. Kraljevski et al. [17] found correlation between DTW distance and the measured Perceptual Evaluation of Speech Quality (PESQ) values for test and received speech in a simulated transmission channel with packet loss. (PESQ [12] is a perceptual objective measure typically used for estimating the transmission channel impact in speech. However, it has been also used for synthesized speech quality assessment [2].)

Algorithm 1 ASL facial expression animations scoring

```
1: function GetDistance($g, c, M, N, c\_dur, anim\_dur$)
2:     G = [g[M1,M2], g[M2,M3], g[M3,M4], g[M4,M5]]
3:     C = [c[T1,T2], c[T2,T3], c[T3,T4], c[T4,T5]]
4:     distance = 0
5:     for ph_g, ph_c in pair (G, C) do
6:         norm_d = DTW(ph_g, ph_c)
7:         distance = distance + norm_d
8:     scale = anim_dur / c_dur
9:     return distance * scale
```

Mana and Pianesi [19] used DTW distance as a quality measure for the quantitative evaluation of synthesized non-sign-language emotional facial expressions in a MPEG-4 compatible avatar. They compared "synthetic" time series of facial markers per frame, with the corresponding "natural" time series performed by a human. While the authors commented that the synthetic animations preferred by DTW appeared (to them) similar to the original human performance, they did not verify that DTW scores related to human judgments of facial expression similarity by conducting a user study (which we have done, as described in section 7).

5.1. Our DTW-based scoring algorithm

Our scoring algorithm assumes that we have:

- A timeline of the words for a "target" ASL animation that we want to generate, where the facial expression has a given duration in milliseconds (**anim_dur**). If we are synthesizing an ASL animation using a pre-built animation lexicon of individual ASL signs, then the duration of these items will affect the overall timeline plan for the target animation to be synthesized. Now, a facial expression must be synthesized.
- A "gold standard" (**g**) motion-capture recording of a human's facial expressions for this ASL sentence (or a very high quality animation of a facial expression which is trusted to be of excellent quality) and the list of five milestones on its timeline (**M1, ..., M5**). Notably, the timeline of when the recorded human performed each word of the sentence will be slightly different than the timeline of the target animation. A video recorded performance of ASL grammatical facial expression can be considered as a multivariate time series, a series of detected MPEG-4 FAPs values in each video frame.
- A "candidate" stream (**c**) of MPEG-4 facial expression parameters that has been synthesized by some software (or perhaps another motion-capture recording) that we wish to evaluate, the list of five milestones on its timeline (**T1, ..., T5**), and its duration in milliseconds (**c_dur**).

Our scoring algorithm initially constructs a list of partial streams for the four phases of the facial expressions g and c based on the intervals between the given five milestones on their timeline (Line 2, Line 3). Then it initializes the total distance between the gold standard and the animated candidate with 0 (Line 4). For each pair of steams of the same phase (Line 5) the algorithm calculates the normalized distance based on Dy-

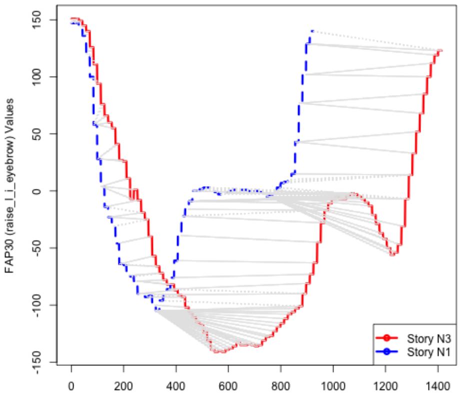


Figure 4: *Example of DTW alignment between the “raise_l_i_eyebrow” values detected in human recordings of two ASL stories containing a Negation facial expression.*

namic Time Warping (Line 6) and adds it to the total distance (Line 7). Since the “candidate” stream and the final animation have different durations, a scaling factor is applied to the distance, based on the stretching or compression of the “candidate” stream (Line 8, Line 9).

To calculate the distance between the two partial streams (Line 6) we used the implementation of multivariate DTW in [4]. It computes a global alignment with minimum distance normalized for path length using Euclidean as a local distance. Computing global alignments means that the time series’ heads and tails are constrained to match each other. We further tuned the algorithm by using the asymmetric step pattern and a SakoeChiba warping window of size 10.

Figure 4 illustrates an example of an alignment for the detected values of MPEG-4 control “raise_l_i_eyebrow” with the Visage SDK [24] during a human’s performance of two ASL stories containing a Negation facial expression (with codenames N3 and N1 in the stimuli collection [9]). The alignment is preformed with the default multivariate implementation of DTW in the R package, *dtw* [4]. The duration of the facial expression in N3 and N1 is 1414 and 924 frames, respectively and their calculated normalized distance was found to be 8.76.

6. Hypotheses

Our goal for this paper is to evaluate our novel, sign-language specific, DTW-based scoring algorithm for facial expressions. One method would be to conduct a study in which human judges estimate similarity scores between face movements in pairs of ASL recordings (and then compare our algorithm to their scores), but we did not find prior published studies in which human judges were able to provide reliable numerical ratings of facial expression similarity between pairs of ASL animations. On the other hand, in several prior studies [8], human participants have been able to answer comprehension questions about ASL animations and indicate whether they noticed particular facial expressions. Thus, we evaluated our DTW algorithm by: (1) selecting a human ASL recording that serves as a gold-standard face performance, (2) using our similarity scoring algorithm to compare this gold-standard to other candidate recordings, and (3) asking human judges to evaluate the comprehensibility of these candidate ASL performances. If we find that our algorithm?s prediction of the similarity between the candidate and the gold-standard correlates with such human-judgments, then we would posit that our algorithm is a useful tool for researchers who are investigating the synthesis of sign-language facial expressions. Thus, we propose the following two hypotheses:

Hypothesis 1: Our scoring algorithm correlates with participants’ implicit understanding of the facial expression, as measured through comprehension questions that probe the participant’s understanding of the information content of the animation.

Hypothesis 2: Our scoring algorithm correlates with participants’ explicit recognition of the facial expression, as measured through a question that asks participants whether they noticed a particular facial expression during the animation.

7. User study

To evaluate our hypotheses, we conducted a user study, where participants viewed animations of short stories in ASL and then answered comprehension and scalar-response questions.

Stimuli. To produce animated stimuli, we selected 6 recordings of a human ASL signer performing ASL stories for each of the 3 categories of ASL grammatical facial expressions (Negation, WH-Question, or Yes-No Question). This is a total of 18 stimuli. We describe our collection of recordings in [9], and the codenames of the selected stories used in this paper were N1-N6, W1-W6, and Y2-Y7, respectively. To obtain the facial expression data that would drive the animations we run Visage Face Tracker [24] on the video recordings of a native ASL signer performing each of the stories. Then we extracted the head position, head orientation, and MPEG-4 FAPs values for the portion of the story where the facial expression occurs.

Next, to generate our stimuli, we rendered an ASL animation of each story in two different versions:

min-distance: Face, head, and torso movements are driven by the recorded performance of the story with the smallest DTW distance from the 5 stories available in the same category. That is, to synthesize an animation of story N1, we used the face and head movements from the story in the set N2-N6 that had the minimum distance from the N1 recording, based on our new scoring algorithm (section 5.1). Notably, stories N2-N6 had different words, but were all Negation stories.

max-distance: Face, head, and torso movements are driven by the recorded performance of the story with the largest DTW distance from the 5 stories available in the same category.

Figure 5 illustrates the two versions of a Yes-No Question story (codename Y3). The video size, resolution, and framerate for all stimuli were identical. The hand movements in each version were identical and were created by native ASL signers using our laboratory’s animation platform [5]. The facial movements were added during the portion of the story where the facial expression of interest should occur; the rest of the story had a static neutral face. The recorded head and facial movements were warped based on the timing of the words in the target animation, as described in section 3. Example stimuli animations from our study are available here: http://latlab.ist.rit.edu/2015slpat.

Experiment Setup. We conducted an evaluation study in which native ASL signers viewed animations of a virtual human character telling a short story in ASL. Each story included instances of one of the facial expressions of interest: Negation, WH-Question, or Yes-No Question. After watching each story

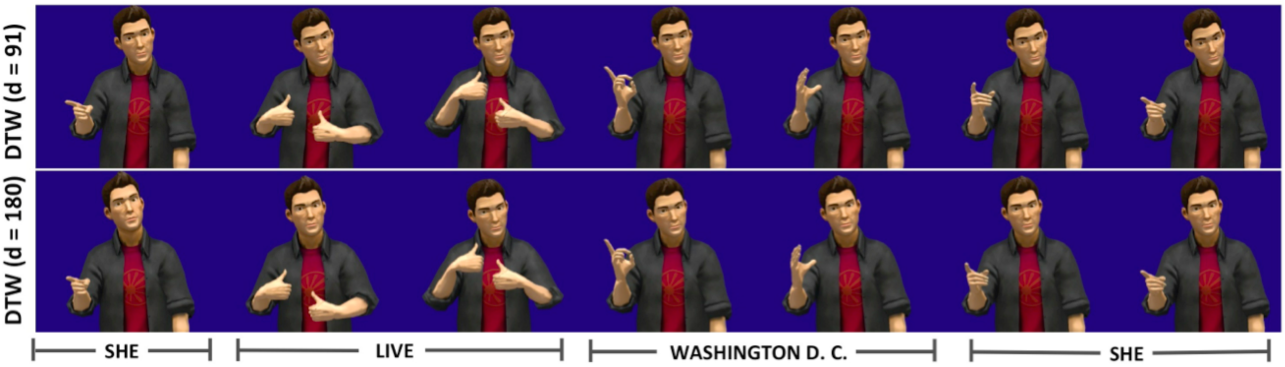


Figure 5: *Screenshots from a min-distance and max-distance version of a Yes-No Question stimulus in the study.*

animation (with facial expressions of one of two types: min-distance or max-distance) one time, participants responded to a "Notice" question (1-to-10 from "yes" to "no" in relation to how much they noticed an emotional, negative, questions, and topic facial expression during the story). Participants were asked to watch the story once more and answer four comprehension questions [9] on a 7-point scale from "definitely no" to "definitely yes." Participants could choose "I'm not sure" instead of answering. As discussed in [15], these stories and comprehension questions were engineered in such a way that the wrong answers to the comprehension questions would indicate that the participants had misunderstood the facial expression displayed [15]. E.g. the comprehension-question responses would indicate whether a participant had noticed a "yes/no question" facial expression or instead had considered the story to be a declarative statement.

At the beginning of the study, participants viewed a sample animation, to familiarize them with the experiment. A native ASL signer conducted all of the experiments in ASL. In prior work [9], we developed methods to ensure that responses given by participants are as ASL-accurate as possible.

Participants. In [10], we discussed the importance of participants being native ASL signers and the study environment being ASL-focused with little English influence; we developed questions to screen for native ASL signers. For this study, ads were posted on New York City Deaf community websites asking potential participants if they had grown up using ASL at home or had attended an ASL- based school as a young child. Of the 18 participants recruited for the study, 15 participants learned ASL prior to age 9, The remaining 3 participants had been using ASL for over 11 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 10 men and 8 women of ages 22-42 (average age 29.8).

8. Results

Our hypotheses considered whether our new scoring algorithm would correlate with participants' implicit understanding of the facial expression (Hypothesis 1) or explicit recognition of the facial expression (Hypothesis 2).

To examine Hypothesis 1, we calculate the correlation between the distance score from the new algorithm and the score from comprehension questions in the user study. We found a significant correlation (Spearman's rho -0.38, $p - value < 0.001$): Hypothesis 1 was supported.

To examine Hypothesis 2, we consider the correlation between the distance score from the new algorithm and the score from the "Notice" question in the study. We found a significant correlation (Spearman's rho -0.33, $p - value < 0.001$): Hypothesis 2 was supported.

9. Conclusions and future work

While we believe that studies with ASL signers are the most conclusive way to evaluate the understandability and naturalness of animations of ASL, our positive results for hypotheses 1 and 2 suggest that sign-language animation researchers could use our new scoring algorithm to evaluate the facial expressions produced by their software. Having a rapid, repeatable method of evaluating the output of facial expression synthesis software is useful for monitoring the development of software, and this evaluation can be performed more frequently than user-based evaluations.

We believe that the time-warping algorithm (section 3) and our scoring algorithm (section 5.1) are a first-attempt at developing an automatic scoring approach, and now that we have observed some moderate though significant correlations in this study, we plan on investigating further variations of these techniques that might prove even more effective. For example, we may investigate the use of Longest Common Subsequence [29] instead of Dynamic Time Warping – or other probabilistic approaches to similarity – and compare them to our findings. We noticed that some of the phases (e. g., M4-M5) of the facial expressions had higher correlations with the participants' scores compared to other phases. This might indicate the need for further tuning of the coefficients for the partial distances calculated on each of the 4 phases.

In future work, we are interested in designing learning-based models for the synthesis of ASL facial expressions, including: topic, rhetorical questions, and emotional affect [13].

10. Acknowledgments

This material is based upon work supported by the National Science Foundation under award number 1506786. We are grateful for assistance from Andy Cocksey, Alexis Heloir, and student researchers, including Christine Singh, Evans Seraphin, Kaushik Pillapakkam, Jennifer Marfino, Fang Yang, and Priscilla Diaz. We would like to thank Miriam Morrow and Jonathan Lamberton for their ASL linguistic expertise and assistance in recruiting participants and conducting studies.

11. References

[1] C. L. BakerShenk, "American Sign Language: A teacher's resource text on grammar and culture," *Gallaudet University*, 1991.

[2] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," *In Proc. European Congress on Acoustics*, pp. 2725–2728, 2005.

[3] S. Gibet, N. Courty, K. Duarte, and T. L. Naour, "The SignCom system for data-driven animation of interactive virtual signers: Methodology and Evaluation," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 1, pp. 6, 2011.

[4] T. Giorgino, "Computing and visualizing dynamic time warping alignments in R: the dtw package," *Journal of statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[5] A. Heloir and M. Kipp, "Real-time animation of interactive agents: Specification and realization," *Applied Artificial Intelligence*, vol. 24, no. 6, pp. 510–529, 2010.

[6] A. Heloir, Q. Nguyen, and M. Kipp, "Signing Avatars: a Feasibility Study," *The Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, Dundee, Scotland, United Kingdom, 2011.

[7] E. Hsu, M. da Silva, and J. Popovi?, "Guided time warping for motion editing," *In Proc. of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 45–52, Eurographics Association, 2007.

[8] M. Huenerfauth and H. Kacorri, "Best practices for conducting evaluations of sign language animation," *In Proc. of the 30th Annual International Technology and Persons with Disabilities Conference (CSUN 2015), Scientific/Research Track*, San Diego, California, USA, 2015.

[9] M. Huenerfauth and H. Kacorri, "Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language," *In Proc. of the Workshop on the Representation and Processing of Signed Languages (LREC 2014)*, Rekjavik, Iceland, 2014.

[10] M. Huenerfauth, L. Zhao, E. Gu, and J. Allbeck, "Evaluation of American sign language generation by native ASL signers," *ACM Trans Access Comput*, vol. 1, no. 1, pp. 1–27, 2008.

[11] ISO/IECIS14496-2Visual, 1999.

[12] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," *ITU Geneva*, 2011.

[13] H. Kacorri, "Models of linguistic facial expressions for American Sign Language animation," *ACM SIGACCESS Accessibility and Computing*, vol. 105, pp. 19–23, 2013.

[14] H. Kacorri, "TR-2015001: A survey and critique of facial expression synthesis in sign language animation," *Computer Science Technical Reports. Paper 403*, 2015.

[15] H. Kacorri, P. Lu, and M. Huenerfauth, "Evaluating facial expressions in American Sign Language animations for accessible online information," *In Proc. of the International Conference on Universal Access in Human-Computer Interaction (UAHCI)*, Las Vegas, NV, USA, 2013.

[16] M. Kipp, A. Heloir, and Q. Nguyen, "Sign language avatars: Animation and comprehensibility," *In Intelligent Virtual Agents*, pp. 113–126, Springer, 2011.

[17] I. Kraljevski, S. Chungurski, Z. Gacovski, and S. Arsenovski, "Perceived speech quality estimation using DTW algorithm," *In 16th TELFOR*, Belgrade, Serbia, 2008.

[18] D. Lillo-Martin, "Aspects of the syntax and acquisition of WH-questions in American Sign Language," *In K. Emmorey & H. Lane (Eds.), The Signs of Language Revisited*, pp. 401–413, Mahwah, NJ: Lawrence Erlbaum, 2000.

[19] N. Mana and F. Pianesi, "HMM-based synthesis of emotional facial expressions during speech in synthetic talking heads," *In Proc. of the 8th international conference on Multimodal Interfaces*, pp. 380–387, ACM, 2006.

[20] R. Mitchell, T. Young, B. Bachleda, and M. Karchmer, "How many people use ASL in the United States? Why estimates need updating," *Sign Lang Studies*, vol. 6, no. 3, pp. 306–335, 2006.

[21] I. Mlakar, and M. Rojc, "Towards ECA?s animation of expressive complex behaviour," *In Analysis of Verbal and Nonverbal Communication and Enactment, The Processing Issues*, pp. 185–198, Springer Berlin Heidelberg, 2011.

[22] C. Neidle, D. Kegl, D. MacLaughlin, B. Bahan, and R. G. Lee, "The syntax of ASL: functional categories and hierarchical structure," *Cambridge: MIT Press*, 2000.

[23] M. Ouhyoung, H. S. Lin, Y. T. Wu, Y. S. Cheng, and D. Seifert, "Unconventional approaches for facial animation and tracking," *In SIGGRAPH Asia*, pp. 24, 2012.

[24] T. Pejsa and I. S. Pandzic, "Architecture of an animation system for human characters," *In Proc. 10th Int?l Conf on Telecommunications (ConTEL)*, pp. 171–176, IEEE, 2009.

[25] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[26] C. Schmidt, O. Koller, H. Ney, T. Hoyoux, and J. Piater, "Enhancing gloss-based corpora with facial features using active appearance models," *Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT)*, 2013.

[27] C. Traxler, "The Stanford achievement test, 9th edition: national norming and performance standards for deaf and hard-of-hearing students," *J Deaf Stud & Deaf Educ*, vol. 5, no. 4, pp. 337–348, 2000.

[28] V. M. Velichko and N. G. Zagoruyko, "Automatic recognition of 200 words," *International Journal of Man-Machine Studies*, vol. 2, no. 3, pp. 223–234, 1970.

[29] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multidimensional time-series," *The VLDB Journal?The International Journal on Very Large Data Bases*, vol. 15, no. 1, pp. 1–20, 2006.

[30] Y. Zhang, Q. Ji, Z. Zhu, and B. Yi, "Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters," *Circuits and Systems for Video Technology*, vol. 18, no. 10, pp. 1383–1396, 2008.

[31] F. Zhou, F. De la Torre, "Canonical time warping for alignment of human behavior," *In NIPS*, pp. 2286–2294.

Qualitative investigation of the display of speech recognition results for communication with deaf people

Agnès Piquard-Kipffer, Odile Mella[1]*, Jérémy Miranda*[1]*, Denis Jouvet*[1]*, Luiza Orosanu*[1]

Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
Inria, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
agnes.piquard@loria.fr, [1]{prenom.nom}@loria.fr

Abstract

Speech technologies provide ways of helping people with hearing loss by improving their autonomy. This study focuses on an application in French language which is developed in the collaborative project RAPSODIE in order to improve communication between a hearing person and a deaf or hard-of-hearing person. Our goal is to investigate different ways of displaying the speech recognition results which takes also into account the reliability of the recognized items. In this qualitative study, 10 persons have been interviewed to find the best way of displaying the speech transcription results. All the participants are deaf with different levels of hearing loss and various modes of communication.

Index Terms: speech recognition, deaf or hard-of-hearing people, compensating for disadvantages, display of speech transcription, French language

1. Introduction

In the world, there are millions of people with hearing loss (http://www.who.int/pbd/deafness/news/Millionslivewithhearingloss.pdf; http://wfdeaf.org). In France over 11% of people suffer from hearing loss which causes several other limitations that are persistent [1]. The sensory problems involve both perceptual, speech, cognitive and social difficulties [2] [3]. The unemployment rate thus varies from 15 to 50% depending on the type of hearing loss.

Deaf adults still have difficulties mastering French language, which is not considered, for some of them, as their native language. Sign language may also not be considered as their native language and has no written modality. The lack of oral interaction is repeated in many situations, even for those adults for whom hearing aids provide correction. In working situations with hearing persons, deaf adults often have to be supported by others [4]. The long term goals of the Rapsodie project (http://erocca.com/rapsodie) are to facilitate the integration of deaf or hard-of-hearing people within a professional context thus aiding their independence, providing them ways of comprehension and communication with automatic speech transcription help.

Our research relates to an embedded system, used in a professional context which could help deaf or hard-of-hearing persons, employees, to interact with a speaking person, customer, without the help of an interpreter. The speech recognition of the customer's utterance is displayed on the screen of the embedded terminal.

The difficulty comes from the fact that speech transcription results contain recognition errors, especially if it is a real time process on a device with limited resources (CPU and memory) and in a noisy environment. As in many real-work conditions, the speech signal is overlapped with parasitic noise, undesired extra speech, or music. These difficulties may impact the understanding processes. There has been many attempts to develop speech recognition appliances but to our knowledge, there is no suitable, validated and currently available screen display of the output of automatic speech recognizer for deaf or hard-of-hearing persons, in terms of size, colors and choice of the written symbols. It is the goal of this first qualitative study, taking account of the previously described technical constraints. We interviewed deaf adults at working age, with different levels of hearing loss and various modalities of communication. Our aim were both to study the feasibility of the project with deaf people of varying profiles, to investigate the more suitable display and to examine which factors the participants consider as being helpful for a better understanding of the speech transcription.

In the following sections, the speech recognition system is described and then the different modalities chosen for displaying the recognition output. Afterwards, we focus on the experimental protocol results conducted with 10 deaf people, discussing how they can be accommodated in order to find the best display of the automatic speech transcription results.

2. Speech transcription system

2.1. Choice of linguistic units

One of the aims of the RAPSODIE project is to realize a portable device embedding a speech recognition system that will help a deaf or hard-of-hearing person to communicate with other people. Due to the limits in memory size and computational power imposed by a portable device, the embedded speech decoder should achieve the best compromise between recognition performance, computational cost, acceptable execution time, and the way of displaying the recognition results for people with hearing loss.

Given a recognition engine, the main constraints relate to the size of the language model and of the lexicon. In this context, we have investigated syllable-based lexicons and hybrid language models [5] [6]. Indeed, the combination of words and syllables allows the recognition of the most frequent words as words and the recognition of the out-of-vocabulary words as sequences of syllables. These investigations led us to use a recognition engine system based on a hybrid trigram statistical language model with a lexicon composed of about 23,000 words and 3,000 syllables. The words and syllables were selected according to their frequency of occurrences in a training corpus of broadcast news, shows

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 36–41,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

and debates from various radio and TV channels. This hybrid model uses only 14 MB of memory space. When applied for the transcription of radio and TV shows (ETAPE [7] development data – 82,000 running words), more than 94% of the output tokens are words, the remaining part (about 6%) corresponds to syllables. An analysis of the results shows that about 70% of the words hypothesized by the decoder are correct (i.e., correctly recognized), and about 60% of the syllables are correct.

Furthermore, the speech recognition engine is built from the PocketSphinx tool [8] and uses as acoustic models, context-dependent phone HMM models with 3 states and 64 Gaussians per state. The acoustic analysis is the standard MFCC (Mel Frequency Cepstral Coefficients) providing 12 static coefficients and the logarithm of the energy per frame with a 10 ms shift. First and second order temporal derivatives are added to the feature vector.

Finally, the recognition engine provides a sequence of words and syllables corresponding to the customer's utterance.

2.2. Use of confidence measure

Speech recognition is not perfect, especially when using an embedded device in a noisy environment. Two types of errors can occur. When the spoken word does not belong to the recognition lexicon (as a word or a sequence of syllables), the recognition engine recognizes it as another lexical unit or as a succession of smaller units acoustically similar to the unknown unit. Furthermore, it can happen that the spoken word is confused with another one when the conditions are different from those used for the training of the acoustic and language models (noisy environment, spontaneous speech, manner of speaking, etc.). Recognition errors will result in additional difficulties for deaf and hard-of-hearing people to understand the spoken sentence.

Confidence measures aims at indicating the reliability of the speech recognition hypotheses. Several approaches for computing confidence measures have been studied in the past [9]. In [10] confidence measures were used to highlight words with low confidence scores in view of helping error correction in a multimodal environment. Along this line, it is always words with low confidence scores that are differentiated, either in a lighter shade for error correction in voicemail transcripts [11], or highlighted for computer assisted speech transcription [12], or displayed with an underlining dependent on the confidence measure [13]. As the confidence measures are not perfect such approaches do not always accelerate the detection and correction of the errors [13]. A few other studies were more concerned with understanding aspects. In [14] the words are displayed with a brightness that depends on their score (kind of confidence measure) in the context of speech playback using time-compression and speech recognition. In all the previous studies, the speech signal was available to the user. This is not the case of [15] which has investigated the understanding of sentences from their speech recognition output only, and investigated how much taking into account the confidence measures in the display can help.

In the current study, we use the confidence measure computed by the speech recognition system to make the result of the recognition easier to understand by deaf users. The speech recognition engine provides a confidence measure for every recognized unit (word and syllable). This measure is based on posterior probability [9]. By comparing the confidence measure to a threshold adjusted on a development corpus, each lexical unit is labeled as "*correctly recognized*" (high confidence score) or "*incorrectly recognized*" (low-confidence score). This characterization (right or wrong) of the words by the recognition system will be displayed on the terminal and different display modes will be proposed for assessment to several deaf persons.

3. On-screen display modalities

3.1. On-screen display modes of the speech recognition results (without using confidence measures)

After the speech recognition process, the recognized words and syllables are displayed on the screen of the portable device. Regardless of the accuracy of the recognition result, it is important to investigate the best way to display this result for deaf and hard-of-hearing people. First, because the result is a mixture of words, and syllables that cannot be written into an orthographical form. Secondly, because for deaf people, orthographic transcription is not necessarily the best way to display the recognition result according to the type of hearing loss and the kind of speech and language training. We decided to study the three following display modes:

- **Orthographic**: the recognized words are written into orthographical form, the syllables are written into pseudo-phonetic form;
- **International Phonetic Alphabet (IPA):** all the recognized words and syllables are written into phonetic form using the International Phonetic Alphabet. Some deaf adults benefited from early hearing and speech intervention which gave them International Phonetic Alphabet knowledge when they learned to read and during speech and language remediation therapy;
- **Pseudo-phonetic**: all the recognized words and syllables are written into a pseudo-phonetic alphabet. Indeed, the phones within the recognized words and syllables are translated into a simple sequence of graphemes using a kind of phonetic spelling. This mode seems appropriate for all the deaf persons who are familiar with French language pronunciation.

An example of a recognition result displayed in these 3 modes is presented Table 1.

Display mode	**Result of the automatic transcription (into words and syllables)**
Orthographic	je voudrais être li vré combien ça kou te
IPA	ʒə vudʁɛ ɛtʁ li vʁe kɔ̃bjɛ̃ sa ku tə
Pseudo-phonetic	je voudré ètr li vré konbyin sa kou te

Table 1: *The different evaluated modes for displaying the result of the recognition of the uttered sentence: "je voudrais être livré, combien ça coûte ?" (I would like it to be home delivered, how much does it cost?).*

	words/syllables tagged as ***incorrect*** are displayed in another color (**red**)	words/syllables tagged as ***correct*** are displayed in **bold**
words tagged as *incorrect* are displayed into **orthographic** mode (syllables are always displayed in pseudo-phonetic mode)	je voudrais être li vré qu'on bien ça kou te	**je voudrais être** li vré qu'on bien **ça** kou te
words tagged as *incorrect* are displayed into **pseudo-phonetic** mode (syllables are always displayed in pseudo-phonetic mode)	je voudrais être l i v r é k on b y in ça k ou t e	**je voudrais être** l i v r é k on b y in **ça** k ou t e

Table 2*: Four screen display modalities to differentiate the words/syllables considered as incorrectly recognized and those considered as correctly recognized by the speech recognition system. Here, the words "qu'", "on" and "bien", and the syllables /li/, /vré/, /kou/, and /te/ are considered as incorrect.*

3.2. On-screen display modalities using the confidence measure

As explained in Section 2.2, the speech recognition system provides an estimation of the recognition correctness for every lexical unit, even if this estimation may be unreliable. Therefore, it is important to find the best way of presenting this information about the word/syllable correctness to the deaf user.

In [15], it has been shown that hearing users infer the correct word from a word considered incorrect by the speech recognition system, more easily when it was written in phonetic form than when it was written in orthographic form. In particular, when several consecutive words were tagged as misrecognized by the system, the hearing user unsuccessfully focused on the word splitting given by the orthographic mode, causing misunderstandings, while the sound sequence of the words was almost free from errors. Instead, the oralization of the sound sequence helped the user to find the right words and thence the meaning of the sentence. Accordingly it seemed to us interesting to study whether these results remain valid for deaf users.

On the one hand, we examined whether it is more favorable to highlight the "*incorrectly recognized*" or the "*correctly recognized*" lexical units.

On the other hand, we distinguished two modes for displaying the "*incorrectly recognized*" words: the orthographic and pseudo-phonetic modes. Note that syllables are always displayed in pseudo-phonetic mode.

Table 2 summarizes the four different display modalities on an example. In the second colon the lexical units tagged as "*incorrect*" are written in a different color (red) than the lexical units tagged as "*correct*" (black). In the third colon, all lexical units are written in blue and the units tagged as "*incorrect*" as written in bold.

4. Methodology

We conducted a qualitative study which goal was to identify the modalities which could help some deaf adults for a better understanding of the speech transcription and to look at how people can use these modalities.

4.1. Participants

The population was selected on the basis of criteria used to define hearing impairment: any disorder of hearing regardless of cause or severity (cf. World Health Organization [11]). As this is a qualitative study using situations created as close as possible to real professional contexts, we selected deaf adults who were working or who were involved in social and cultural associations, thus well integrated socially despite their communication difficulties. A preliminary selection was made to ensure a functional literacy level, as they would have to read the written transcription of speech recognition.

- Our heterogeneous population, consisted of 10 deaf persons, 4 women and 6 men; from 25 years old to 63 years old, the average age being 39 years,
- 4 persons presenting profound hearing loss, 4 severe hearing loss, 2 severe-moderate loss. The time of acquisition of their hearing loss varied from the first few days, to months or years of life. Most of causes were listed as unknown.

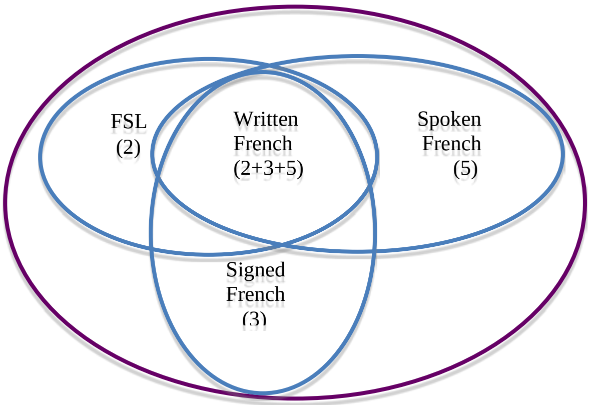


Figure 1: *Distribution of the 10 participants according to their main mode of communication.*

- For some of them, their mother tongue was French or French Sign Language and for some others, neither French nor French Sign Language were considered as their native language. Nine persons regularly used hearing aids to obtain as much as possible of their acoustic information. Various modes of communication were used by the deaf persons: French oral and written Language; French oral Language and French cued-speech (LPC: manual cues to supplement speech input); French written Language; French Sign Language (FSL); fingerspelling (dactylology); "Signed French" (français signé) combining the use of the FSL signs ordered according to the French language linear syntax and fingerspelling. Figure 1 shows the distribution of the 10 participants according to their main mode of communication. The larger outer oval includes the whole set of participants; in each of the three inner ovals are the deaf persons with their specific mode of communication, all of them using written French.

4.2. Tasks and Procedure

Our study was conducted in two phases. For every participant, each phase consisted of several 2-hour sessions including tests and interviews.

Before these two phases, the level of literacy was tested prior to commencing trial. The deaf person had to read a 10-line text describing communication situations which may be encountered in everyday life and in the particular situation: "do-it-yourself" shop. The deaf person has to understand the role he would play: an employee, while the hearing person (the interviewer) would play that of the customer, either at the cash-desk or in the store. In order to verify his comprehension, the participant had to reformulate the text, with his own communication tools.

4.2.1. First phase: Tests and interviews

The goal of the first test was to find the best way of displaying the speech transcription results among the orthographic, IPA and pseudo-phonetic display modes (cf. section 3.1). The confidence measures were not used at this stage.

In this first phase, the participants were required to read and to understand the transcriptions of 10 uttered sentences, the transcriptions were provided by the speech recognition system always in the context of the previous described scenario (do-it-yourself shop).

We elaborated every sentence according to lexical, syntactical and semantic criteria. The main lexical fields were the one of the do-it-yourself and that of the request for commercial information. Syntactically, every sentence was comprised of one or several clauses (constituent of the sentence made up of a subject and a verbal group). The sentences were coherent, reasonably long in order to be as well understood as possible. The average length of the sentences was 11.35 words (minimum: 5 words, maximum: 22 words). Every sentence contained a verb. Declarative, imperative, exclamatory sentences were included with a majority of interrogative sentences, as the test situation was as close as possible to a real situation when the client request information.

The participants were seen individually in a quiet room. They could not be helped by the sound, they had to read the speech transcription of the sentence and try to interpret it and to rephrase it so that the interviewer could check their understanding.

Their answers were not been timed. Rather, each person was interviewed in order to identify the helping points in his/her comprehension processes, sentence by sentence, knowing that speech transcription is not perfect and have no punctuation mark which could indicate the declarative, interrogative, exclamatory and imperative sentences.

We made aware deaf persons of the presence of recognition errors in the transcription system for several reasons:

- So that the deaf adults could not consider the present recognition system as a final perfect tool, as it is still in evolution,
- The correct recognized words and the presence of errors were both the base of discussion with the deaf persons who indicated the points in the display which aided their comprehension.

4.2.2. First phase: Results

The IPA display mode was by far the most difficult to apprehend, therefore none of the participants have indicated it as helpful, this coding requiring special learning. Table 3 shows their preferences. Not even the two deaf persons who still used it in speech remediation therapy found it helpful in such a context. For both familiar and unfamiliar users, reading a whole sentence in IPA required too much time and cognitive resources. Therefore, this display mode was abandoned for both words and syllables.

The pseudo-phonetic display mode was preferred by one participant for both words and syllables. This person indicated an order of usage preference: firstly the pseudo-phonetic mode and then the orthographic display mode, suggesting that the terminal screen could display those two options so that the deaf person could choose the more helpful one.

Display mode	**Preference of participants** (N=10)
Orthographic	9
IPA	0
Pseudo-phonetic	1

Table 3*: The display mode preferred by the participants.*

The orthographic display mode was preferred by almost all participants: nine out of ten. They have all further specified that this mode was aiding (first preference) except in the case of speech recognition errors. In fact, in case of orthographical error, for example for a word pronounced [samədi] corresponding to the word "samedi" ("Saturday") but transcribed as "ça me dit" ("it's tempting"), these deaf persons reported their difficulties to comprehend the whole sentence. The transcribed sentence is segmented differently, including several words instead of one, coming from other grammatical categories and lexical fields: word and time semantic field *versus* sentence and emotion semantic field. In such a case, for the five participants who were more familiar with French language phonology, it was easier to read words into pseudo-phonetic mode, and to infer semantic signification from pronunciation.

Moreover, all the participants considered that displaying the pauses detected by the speech recognizer was helpful.

4.2.3. Second phase: Test and Interviews

The goal of this second phase was to find the best way of displaying the additional information provided by the speech recognizer concerning the correctness of the recognized lexical unit using confidence measure. For that purpose, the four modalities described in the section 3.2 were evaluated. As it is shown in Table 2, in the case of highlighting the "*incorrectly recognized*" lexical units, we chose to display them in another color (red); in the case of highlighting the "*correctly* *recognized*" lexical units, we chose to display them in the same color but in bold.

Two experiments were conducted. Firstly, we used an "*oracle*" confidence measure: the lexical units tagged as "*incorrectly recognized*" were actually the units misrecognized by the speech decoder and, respectively, the lexical units tagged as "*correctly recognized*" were actually the units well recognized by the speech decoder. Secondly, we used the confidence measures computed by the speech recognizer to tag the recognized units.

The same procedure as the one conducted in the first phase was used here.

4.2.4. *Second phase: Results*

Regardless the way in which the transcribed units were tagged (oracle or from real confidence measures), the preferences of the participants were the same. The modality highlighting the "*correctly recognized*" lexical units in bold blue was preferred by all participants. They reported that their major attention was thus focused on words characterized as right (even if, in some cases, they are actually wrong). That was helping them for direct access to understanding. Table 4 summarizes the choices of the deaf persons.

Within this modality, the display into pseudo-phonetic of the words tagged as "*incorrect*" was preferred by a majority of participants, 8 persons, for the reasons previously detailed in section 4.2.2. They also explained that compared to the IPA, this system was using a simple coding scheme. They also reported that this display mode required the use of the context, and time to adapt. Indeed, this system leads to an indirect access to meaning, implying knowledge of phonology, breaking words into syllables in order to « sound out » with the aim of understanding. They also reported that any absence of a pseudo-phoneme made the task very difficult.

	words/syllables tagged as ***correct*** are displayed in **bold**	Preference of the participants (N=10)
words tagged as *incorrect* are displayed into **orthographic** mode	**je voudrais être** li vré qu'on bien **ça** kou te	2
words tagged as i*ncorrect* are displayed into **pseudo-phonetic** mode	**je voudrais être** l i v r é k on b y in **ça** k ou t e	8

Table 4*: The display modalities preferred by the participants.*

The display into orthographic mode of the words tagged as "incorrectly recognized" was preferred by two persons who therefore indicated weak points of this display mode. The words characterized as "incorrect" by the recognition system could place them in serious difficulties; those words could be in contradiction with the signification of the remaining part of the sentence (cf. 4.2.2). Nevertheless, they didn't feel familiar enough with French phonology to dare using the pseudo-phonetic mode.

5. Discussion and conclusion

In the context of improving communication between a hearing person and a deaf person, when displaying on an embedded device the results of an automatic speech transcription system, highlighting in bold the words considered as "*correctly recognized*" rather than the words considered as "*incorrectly recognized*" is more helpful. All the participants stressed that knowing the context and searching for keywords are essential steps to build their capacity of understanding. Highlighting the words considered as "*correctly recognized"* enables them to construct inferences, and to gain confidence, provided that there is an adequate number of key elements clearly identified.

The display into pseudo-phonetic of the words tagged as "*incorrectly recognized*" was preferred by a majority of participants (8), those persons were more familiar which French language including phonology. These results are similar to those showed from a previous study undertaken among a hearing population [15].

However, they explained that a training phase would be necessary to get more familiar with pseudo-phonetic reading. It could improve their understanding and in the long term facilitate the communication with speaking persons.

The other two persons who preferred the words tagged as "*incorrect*" displayed into orthographic mode were those who mainly use French Sign Language. Unfortunately, for them this display mode is not aiding enough in case of errors. Their comprehension processes cannot be supported by enough reliable words. They have to guess with many risks of misunderstanding and discouragement.

At a general level, the interviews showed that it was difficult for all the participants to stay aware of the fact that the cues based on computed confidence measures are not fully reliable. This was expressly mentioned when the participants could read the sentence with sufficient understanding, considering it as appropriate to the particular context. It was difficult for them to assess whether the information was to be trusted. The same difficulties have been observed in [13], in an experiment in which hearing people dictated a text and then had to detect the errors made by the speech recognition.

Our preliminary qualitative study was conducted in the worst conditions as the participants had only the written sentences with no oral pronunciation. They could not rely on their hearing aids nor lips reading to help them and the context information was limited. The tests were conducted in a quiet neutral room and not in a "do-it-yourself" shop. Thus, the participants could not be helped by the context of the shop (customer, special department, visual cues). As, in those experiments, no punctuation was indicated in the speech transcriptions, the deaf persons had difficulties to differentiate interrogative sentences from declarative ones.

Nevertheless, all the participants showed their interest for such a system and thought that it could be more helpful with the help of context. Further experimentations will be conducted to investigate the efficiency of this system compared to or combined with other communication means used by deaf and hard-of-hearing persons.

6. Acknowledgements

The work presented in this article is part of the RAPSODIE project, and has received support from the "Conseil Régional de Lorraine" and from the "Région Lorraine" (FEDER) (http://erocca.com/rapsodie).

We would like to thank The "Institut des Sourds de La Malgrange", "Espoir Lorrain", an independent association for the deafened and hard-of-hearing people, the deaf persons

interviewed, Marie-Madeleine Dutel, Jane and Vivien Dasset for their help.

7. References

[1] L. Haeusler, T. De Laval, C. Millot, “Etude quantitative sur le handicap auditif à partir de l’enquête *Handicap-Santé*" , *DREES* Direction de la recherche, des études, de l’évaluation et des statistiques, *Document de travail, Série Etudes et Recherches, 131*, août 2014.

[2] H. R. Myklebust, “The psychology of deafness”, New York: Grune & Stratton, 1964.

[3] A. Dumont, “Orthophonie et Surdité. Communiquer, comprendre, parler”, Masson, collection orthophonie, 2008.

[4] D. R. Calvert & S.R. Silverman, “Speech and deafness”, *A.G. Bell Association for the Deaf, Washington, D.C.* 1975.

[5] L. Orosanu and D. Jouvet, “Comparison of approaches for an efficient phonetic decoding,” *in INTERSPEECH 2013. Lyon, France*, 2013.

[6] L. Orosanu and D. Jouvet, “Hybrid language model for speech transcription”, *in INTERSPEECH 2014, Singapour*, 2014.

[7] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, “The ETAPE corpus for the evaluation of speech-based tv content processing in the French language”, *in Language Resources and Evaluation (LREC’12)*, 2012.

[8] D. Huggins-daines, M. Kumar, A. Chan, A.-W. Black, M. Ravishankar, and A. Rudnicky, “Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices,” *in ICASSP, Toulouse, France*, 2006.

[9] H. Jiang, “Confidence measures for speech recognition: A survey,” *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[10] B. Suhm, B. Myers, A. Waibel. “Multimodal error correction for speech user interfaces”, *ACM Transactions on Computer Human Interaction (TOCHI)*, 8(1), 60-98, 2001.

[11] M. Burke, B. Amento, P. Isenhour, “Error correction of voicemail transcripts in scanmail”, in *Proc. SIGCHI conference on Human Factors in Computing Systems*, pp. 339-348, ACM, 2006.

[12] S. Luz, M. Masoodian, B. Rogers, C. Deering, “Interface design strategies for computer-assisted speech transcription”, in *Proc. 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat*, pp. 203-210, ACM, 2008.

[13] K. Vertanen, P.O. Kristensson, “On the benefits of confidence visualization in speech recognition”, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1497-1500, ACM, 2008.

[14] S. Vemuri, P. DeCamp, W. Bender, C. Schmandt, “Improving speech playback using time-compression and speech recognition”, in *Proc. SIGCHI conference on Human factors in computing systems*, pp. 295-302, ACM, 2004.

[15] J. Razik., O. Mella, D. Fohr, J.-P. Haton, “Frame-Synchronous and Local Confidence Measures for Automatic Speech recognition”, *International Journal of Pattern Recognition and Artificial Intelligence*, 2011.

[16] CIM-10, “Classification internationale des maladies. Chapitre VIII: maladie de l’oreille et de l'apophyse mastoïde. H90. Surdité de transmission et neurosensorielle”. 10eme révision. Paris : Masson, 1993.

Predicting disordered speech comprehensibility from Goodness of Pronunciation scores

Lionel Fontan[1], *Thomas Pellegrini*[1], *Julia Olcoz*[2], *Alberto Abad*[3,4]

[1]Université de Toulouse; UPS; IRIT; Toulouse, France
[2]ViVoLAB - Voice Input Voice Output Laboratory; I3A; Universidad de Zaragoza, Zaragoza, Spain
[3]L2F - Spoken Language Systems Laboratory; INESC-ID; Lisbon, Portugal
[4]IST - Intituto Superior Técnico, Universidade de Lisboa, Portugal
{lionel.fontan, thomas.pellegrini}@irit.fr, jolcoz@unizar.es, alberto.abad@l2f.inesc-id.pt

Abstract

Speech production assessment in disordered speech relies on tests such as intelligibility and/or comprehensibility tests. These tests are subjective and time-consuming for both the patients and the practitioners. In this paper, we report on the use of automatically-derived pronunciation scores to predict comprehensibility ratings, on a pilot development corpus comprised of 120 utterances recorded by 12 speakers with distinct pathologies. We found high correlation values (0.81) between Goodness Of Pronunciation (GOP) scores and comprehensibility ratings. We compare the use of a baseline implementation of the GOP algorithm with a variant called forced-GOP, which showed better results. A linear regression model allowed to predict comprehensibility scores with a 20.9% relative error, compared to the reference scores given by two expert judges. A correlation value of 0.74 was obtained between both the manual and the predicted scores. Most of the prediction errors concern the speakers who have the most extreme ratings (the lowest or the largest values), showing that the predicted score range was globally more limited than the one of the manual scores due to the simplicity of the model.

Index Terms: pronunciation assessment, Goodness of Pronunciation, disordered speech, comprehensibility

1. Introduction

The assessment of speech production abilities in motor speech disorders relies almost exclusively on subjective tests such as intelligibility tests. These tests have two main disadvantages. They are very time-consuming and often imply subjective judgments: speakers read lists of words or sentences while one or several judge(s) evaluate their production. Within this framework automatic methods for speakers evaluation appear as practical alternatives. Recent advances in Automatic Speech Recognition (ASR) – especially in the field of Computer-Assisted Language Learning (CALL) – have contributed to develop techniques that may be of great interest for this purpose.

ASR techniques developed for the assessment of foreign language learners' pronunciation skills focused both on the segmental and the suprasegmental levels, giving birth to two research fields respectively called *individual error detection* and *overall pronunciation assessment* [1]. For individual error detection (i.e., automatic detection of mispronounced phones), two kinds of methods are used:

- methods based on the comparison of target phone models and learners' phone models (e.g. *nonnativeness* [2] or scores derived from classification methods such as linear discriminant analysis and alike [3]);
- methods independent of the learner's native language, such as raw recognition scores [4], or Goodness of Pronunciation scores (GOP [5, 6]).

Since the latter methods do not rely on any assumption concerning the errors possibly made by the speakers, their relevance may not be limited to the field of CALL. For example, GOP scores can be calculated to get an idea on how confident the ASR system is about each phone identity. In a previous research work [7], GOP scores were compared to perceptual analysis results in order to detect mispronounced phonemes in individuals with unilateral facial palsy (UFP). The algorithm was found to be effective: it detected 49.6% of mispronunciations (CR rate) and 84.6% of correct pronunciations. In [8] a preliminary test was conducted in order to study the relationship between mean GOP scores at sentence-level and subjective comprehensibility. Results were encouraging as highly significant correlations were observed, with absolute Pearson's coefficients ranging from .68 to .79.

However, several questions remain concerning this last study. First, only the baseline implementation of the GOP algorithm was used. Recent algorithm refinements for CALL applications suggest that the accuracy of GOP results can be greatly improved, as in Forced-aligned GOP measurements (F-GOP [9]). Moreover, the ability of GOP scores to predict comprehensibility judgments or measures was not assessed since the number of speakers was too limited. As a consequence the aim of the present work is twofold: 1) comparing the efficiency of GOP vs. F-GOP scores when dealing with disordered speech and 2) extending the number of speakers so as to test the ability of GOP measures to actually predict comprehensibility.

2. GOP algorithms

The purpose of the GOP algorithm is to automatically provide pronunciation scores at segmental level, that is one score per phone realization. The larger the score, the larger the difference between a phone realization and the corresponding phone model. In other words, large scores indicate potential mispronunciations. In this work, we used two different implementations: the original "baseline" one [5, 6], and a variant called Forced-aligned GOP (F-GOP) [9].

The baseline algorithm can be decomposed into three steps: 1) forced phone alignment phase, 2) free phone recognition phase and 3) score computation as the difference between the

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 42–46,
Dresden, Germany, 11 September, 2015.

Table 1: *Mean GOP values, reaction time and comprehensibility scores for 6 speakers. AP: Patients suffering from structural (anatomic) disorders, NP: Patients suffering from neurological disorders*

Speaker	Mean GOP value	Mean F-GOP value	Mean Reaction Time to oral commands (s)	Mean comprehensibility score
AP1	1.60 (0.56)	0.81 (0.36)	4.11 (0.77)	5.65 (0.45)
NP1	2.32 (0.66)	1.11 (0.38)	4.63 (1.08)	5.30 (0.40)
NP2	2.54 (0.48)	1.42 (0.77)	5.54 (1.17)	4.70 (0.40)
AP2	2.86 (0.71)	1.99 (0.58)	5.50 (1.20)	4.05 (0.45)
AP3	3.67 (0.46)	2.50 (0.68)	7.51 (1.15)	4.25 (0.35)
AP4	4.15 (0.67)	4.01 (1.18)	9.64 (2.56)	1.65 (0.25)

log-likelihoods of the two preceding phases for each forced-aligned phone. The forced alignment phase is intended to provide the ASR system with the orthographic transcription of the input sentence along with a pronunciation lexicon. It consists of forcing the system to align the speech signal with an expected phone sequence. On the contrary, free phone recognition determines the most likely phone sequence matching the audio input without constraint (free phone loop recognition). GOP scores typically range from zero (perfect match) to values up to 10. Higher values often indicate that the aligning step failed for some reason and scores are meaningless in this case. In order to decide whether a phone was mispronounced ("rejected") or not ("accepted"), phone-dependent thresholds can be determined on a development set. In this work, our goal was not to detect individual mispronunciations but rather to compute average GOP scores per utterance in order to correlate them with comprehensibility scores given by human judges at utterance-level.

The forced-aligned GOP version is exactly the same as the baseline one with the only difference that the phone boundaries found during forced alignment constrain the free phone recognition phase. For each aligned phone, a single phone is recognized. In [9], better correlations between GOP and manual scores were found with F-GOP than with baseline GOP in the context of a CALL experiment. Indeed, F-GOP removes the issues of comparing a single aligned phone with potentially several phones recognized within the same time interval.

3. Main objective and methodology

This study aims at verifying the ability of GOP measures to predict disordered speech comprehensibility. To this end, 12 pathological speakers were recorded. In a first experiment, these recordings were split in two subsets, each consisting of the sentences (imperative commands) recorded by 6 speakers: a development corpus and a test corpus (section 4). Reference comprehensibility scores, presented in section 5, were obtained a) by asking 24 listeners to react to the sentences using software created for this purpose and b) by asking two trained speech pathologists to evaluate each sentence comprehensibility on a 7-points rating scale. Automatic measures found in GOP experiments (section 6) are compared so as to establish a predictive model of speakers' comprehensibility. This model is finally used to predict speech pathologists' comprehensibility judgments in 6 other patients (section 7). Since data from 6 speakers constitute a very small dataset with 60 utterances only, we also report prediction results in a cross-validation setup.

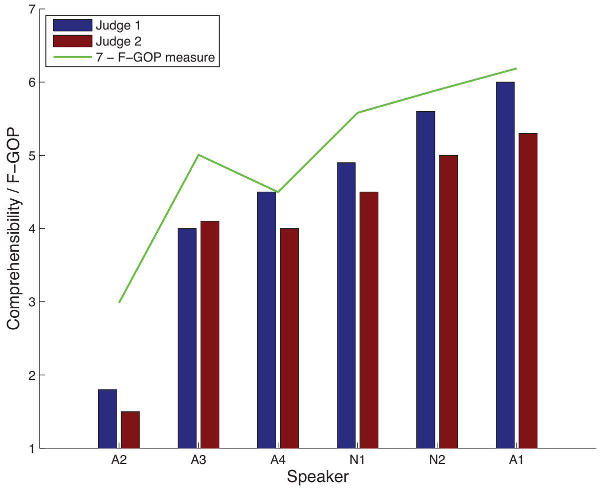


Figure 1: *Comprehensibility judgments as a function of mean F-GOP scores. For a better clarity, F-GOP scores have been scaled following the equation:* $y = 7 - FGOP$.

4. Corpus description

Speech stimuli were recorded from three female and nine male patients. Patients were aged from 33 to 70 years old (mean $= 55$). Four patients suffered from speaking issues due to neurological disorders (spasmodic dysphonia, parkinsonian dysarthria (2) and Huntington's disease) and eight patients had troubles related to anatomic disorders: seven patients suffered from sequelae consecutive to oropharyngeal cancer surgery (among which two total laryngectomees) and one patient had dysphonia. The 12 patients were divided into two groups, both consisting in two patients suffering from neurological speech disorders and four patients suffering from anatomic speech disorders.

Each patient recorded 10 oral commands (sentences) among a hundred different ones, asking to move entities (animals or objects), such as "Mettez l'ours à gauche du kangourou" (*Move the bear to the left of the kangaroo*), or "Mettez le lion sous la banane" (*Move the lion below the banana*). All the commands had the same syntactic form.

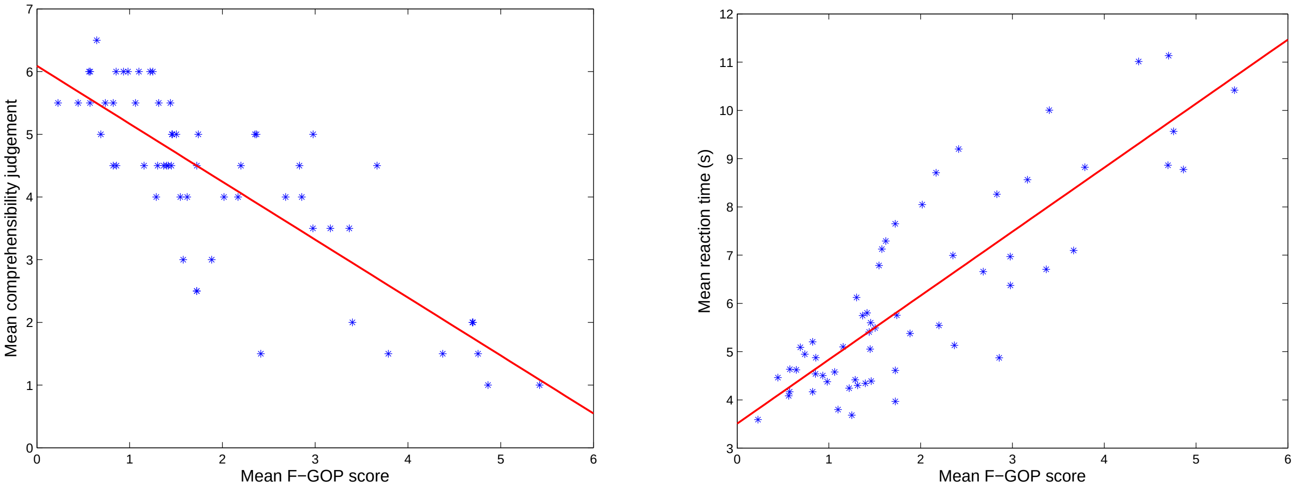


Figure 2: *Left: Mean sentence comprehensibility as a function of F-GOP scores. Ratings range from 1 (very difficult to understand) to 7 (very easy to understand). The red line is regression fit of equation* $y = -.92 *$ *F-GOP* $+ 6.09$*, Right: Mean reaction times to oral commands as a function of F-GOP. The red line represents the regression fit of equation* $y = 1.33 *$ *F-GOP* $+ 3.51$*,*

5. Comprehensibility measures

5.1. Subjective judgments of speech comprehensibility

Two speech pathologists judged each sentence on a 7-points comprehensibility scale, ranging from 1 – *very hard to understand* up to 7 – *very easy to understand*. Both speech pathologists had more than 10 years of experience in listening and evaluating disordered speech. A Kendall tau-b rank correlation was computed so as to check the inter-rater agreement; a highly significant and strong correlation between the two rater scores was found ($t = .73; p < .001$). Finally, mean subjective comprehensibility scores were calculated for each sentence by taking into account the two speech pathologists' grades.

5.2. Behavioral scores: reaction times to oral commands

Behavioral scores were collected for the 60 sentences forming the development corpus. For this purpose 24 listeners responded to the oral commands on a software created for recording their answers and reaction times [10]. For each command six images were displayed on a screen and listeners were asked to move the target image as demanded. As soon as the listener selected an image in order to move it, reaction time (RT) was collected. Keeping as an example the sentence asking to move the bear to the left of the kangaroo, RT was the time elapsed between the beginning of sentence play and the time at which the listener clicked on the image representing a bear. Only cases in which the listeners selected the right target image were considered. Listeners had a mean age of 32.5 years old (SD $= 13.4$) and benefited from various years of experience in listening to disordered speech (mean $= 7.8$; SD $= 11.4$). However, these two variables were found to have a comparable strength and opposite influence on RT [11]; consequently RT have not been weighted as a function of listeners' age and years of experience with disordered speech. Only mean RT for each sentence was taken into account.

Table 2: *Pearson correlation coefficients between automatic scores and comprehensibility measures*

Variables	Correlation
GOP * Comprehensibility ratings	-.684**
F-GOP * Comprehensibility ratings	-.808**
GOP * Reaction times	.786**
F-GOP * Reaction times	.844**

** Correlation is significant at the .001 level (2-tailed)

6. Relationship between GOP scores and speakers' comprehensibility

This section is solely concerned with data issued from the development corpus. Results concerning the prediction of comprehensibility scores from the test corpus will be presented in section 7.

6.1. ASR system setup

This work was carried out with HTK [12]. The acoustic models are three-state left-to-right HMMs with 32 Gaussian mixture components trained on the ESTER corpus [13]. As they have been found to be more suitable for CALL applications [14], context-independent acoustic models (39 monophones) were used.

6.2. Results

6.2.1. Mean scores

Table 1 presents mean and standard deviations of GOP and F-GOP values as well as mean comprehensibility scores for each speaker of the development corpus. Mean RT tend to increase with mean GOP and F-GOP scores, whereas mean comprehensibility appears to decrease as a function of GOP and F-GOP. This suggests that the highest GOP and F-GOP scores are associated with the least comprehensible speakers, and vice versa.

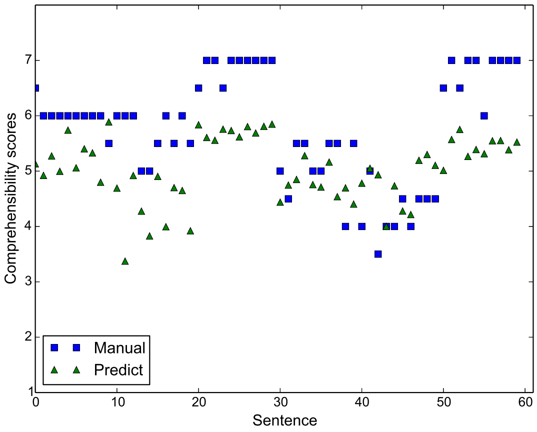


Figure 3: *Manual and predicted comprehensibility scores for each sentence of the test group (6 speakers). Each speaker recorded 10 sentences, so sentences from 0 to 9 on the X-axis correspond to speaker A5, from 10 to 19 to speaker A6.*

6.2.2. Correlation between GOP scores and comprehensibility judgments

Pearson product-moment correlation calculations were computed to study the relationship between GOP/F-GOP scores and comprehensibility measures. Results show a weaker correlation with GOP scores ($r = -.684; p < .001$) than with F-GOP scores ($r = -.808; p < .001$). Both correlations are negative, showing that comprehensibility judgments tend to increase as GOP scores decrease. To illustrate this, comprehensibility and mean F-GOP scores are represented in Figure 1. The correlation plot for all the sentences' F-GOP scores is shown on the left-hand side part of Figure 2.

6.2.3. Correlation between GOP scores and reaction times

For both GOP and F-GOP scores, Pearson product-moment correlation calculations indicate a strong and highly significant relationship with reaction times to oral commands. A stronger correlation is found with F-GOP scores ($r = .844$; $p < .001$) than with GOP scores ($r = .786; p < .001$). The correlation plot for F-GOP scores is shown on the right-hand side part of figure 2. All correlation coefficients found for GOP scores and F-GOP scores are presented in table 2.

7. Prediction of speakers' comprehensibility

As F-GOP are strongly correlated to the patients' comprehensibility scores, a second part of the present work focused on the ability of F-GOP scores to predict speakers' comprehensibility.

7.1. Separate test set

To this end, a first experiment consisted in estimating comprehensibility ratings for 6 "test" patients, different from the ones for which we reported results so far, with the help of the linear regression model previously described in Section 6.2.2. Predicted scores were compared to the mean comprehensibility ratings given by the two speech pathologists. In figure 3, manual and predicted comprehensibility scores are illustrated per sentence. Even if manual and predicted comprehensibility scores seem to follow the same tendencies ($r = .59$) predicted scores appear to be globally lower than manual scores, with a 16.3 % relative mean difference between both scores. This is mainly due to the fact that the 60 utterances were not sufficient to estimate a model.

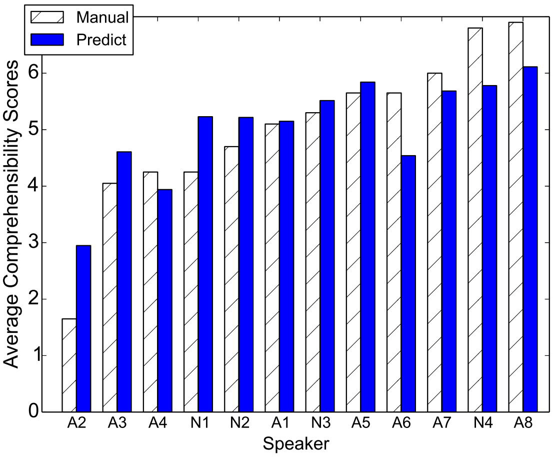


Figure 4: *Mean manual and predicted scores for all the 12 speakers, obtained in the LOSO-CV setup.*

7.2. LOSO-CV setup

In order to obtain sounder results, we repeated the experiment in a Leave-One-Speaker-Out Cross-Validation (LOSO-CV) fashion that allows to use more data to estimate the regression parameters. It corresponds to using data from 11 speakers (110 utterances) for the estimation of the regression parameters (slope and intercept), and to make predictions for the 12^{th} speaker that was left out. This process is repeated for each of the 12 speakers. A global Pearson correlation value of $r = .74$ was obtained, a much larger value than the preceding one. The relative mean difference is higher, though, with a value of 20.9 %. This is probably due to the fact that we make predictions for 12 speakers, twice as many speakers as in the preceding setup. Figure 4 shows a comparison of manual and predicted comprehensibility scores for all the 12 speakers. It shows that the dynamic range of the regression model is too limited: small and large scores are not predicted as accurately as medium scores.

8. Conclusions

The first noticeable result from this study is that a strong and highly significant relationship was found between GOP-derived scores and comprehensibility measures in the particular case of disordered speech. More precisely, the strongest correlations were found with F-GOP measures [9], which presented better results than conventional GOP scores [5]. This observation tends to present F-GOP scores as more closely related to speech production performance, as it was also observed in [9] and [14] in the application domain from which these two algorithms originate – namely Computer-Assisted Language Learning (CALL).

These encouraging results represented a strong motivation for studying the ability of F-GOP scores to predict disordered speech comprehensibility, which was done in the second part of this work. In a first score prediction experiment, data from 6 speakers (60 utterances) were used to estimate a simple linear regression model, and 60 comprehensibility automatic scores were predicted with this model on the remaining utterances from the 6 left-out speakers. A relative mean error of 16.3% was found, together with a low correlation value of 0.59, when comparing the automatic and the manual scores. These results were not conclusive mainly because of the small size of the subset used to estimate the regression parameters. The same prediction experiment but in a cross-validation setup was more satisfying since a 0.79 correlation value was obtained. Nevertheless, the range of the automatic scores still was too small to correctly predict scores from speakers with low and large comprehensibility ratings.

As a response to these observations, future work will be devoted to the enlargement of the pathological speech data, by collecting speech representative of a wide variety of speech disorders. More complex regression models, such as Bayesian models, will be interesting to test. Such models allow to introduce *a priori* information that may help in handling potential differences in model fits that may be seen for different groups of pathological speakers. Adding features characterizing suprasegmental aspects such as speech rate and pitch range, for instance, will also be worth testing.

9. Acknowledgments

The authors would like to acknowledge the Voice and Deglutition Unit (UVD) of the Hospital Larrey at Toulouse, with special regards to Virginie Woisard, Julie Bordenave and Astrid Ros, who participated in collecting the pathological speech data used in this study.

10. References

[1] M. Eskenazi, “An overview of spoken language technology for education,” *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] G. Kawai and K. Hirose, “A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training,” in *Proc. Internat. Conf. on Spoken Language Processing – ICSLP-1998*, 1998, pp. 1823–1826.

[3] H. Strik, K. P. Truong, F. de Wet, and C. Cucchiarini, “Comparing classifiers for pronunciation error detection.” in *Proc. Interspeech 2007*, 2007, pp. 1837–1840.

[4] B. Sevenster, G. d. Krom, and G. Bloothooft, “Evaluation and training of second-language learners’ pronunciation using phoneme-based HMMs,” in *Proc. STiLL*, Marholmen, 1998, pp. 91–94.

[5] S. Witt, “Use of Speech Recognition in Computer-Assisted Language Learning,” PhD Thesis, University of Cambridge, Dept. of Engineering, 1999.

[6] S. Witt and S. Young, “Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning,” vol. 30, pp. 95–108, 2000.

[7] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, “The goodness of pronunciation algorithm applied to disordered speech,” in *Proc. Interspeech 2014*, 2014, pp. 1463–1467.

[8] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, C. Alazard-Guiu, M. Robert, and P. Gatignol, “Automatic assessment of speech capability loss in disordered speech,” *ACM Transactions on Accessible Computing*, vol. 6:3, May 2015.

[9] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, “Analysis and Utilization of MLLR Speaker Adaptation Technique for Learners Pronunciation Evaluation,” in *Proc. Interspeech 2009*, 2009, pp. 608–611.

[10] L. Fontan, P. Gaillard, and V. Woisard, “Comprendre et agir : les tests pragmatiques de comprhension de la parole et elokanz,” in *La voix et la parole perturbes*, R. Sock, B. Vaxelaire, and C. Fauth, Eds. Mons: CIPA, 2013, pp. 131–144.

[11] L. Fontan, “De la mesure de l’intelligibilité à l’évaluation de la compréhension de la parole pathologique en situation de communication,” PhD thesis, Université de Toulouse, 2012.

[12] S. Young and S. Young, “The HTK Hidden Markov Model Toolkit: Design and Philosophy,” *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[13] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, “The ESTER phase II evaluation campaign for the rich transcription of French broadcast news,” in *Proc. Interspeech 2005*, 2005, pp. 1149–1152.

[14] T. Kawahara and N. Minematsu, *Tutorial on CALL Systems at Interspeech*, Portland, 2012.

Recognizing Dysarthric Speech due to Amyotrophic Lateral Sclerosis with Across-Speaker Articulatory Normalization

Seongjun Hahm[1], *Daragh Heitzman*[3], *Jun Wang*[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]MDA/ALS Center, Texas Neurology, Dallas, Texas, United States
{seongjun.hahm, wangjun}@utdallas.edu; dheitzman@texasneurology.com

Abstract

Recent dysarthric speech recognition studies using mixed data from a collection of neurological diseases suggested articulatory data can help to improve the speech recognition performance. This project was specifically designed for the speaker-independent recognition of dysarthric speech due to amyotrophic lateral sclerosis (ALS) using articulatory data. In this paper, we investigated three across-speaker normalization approaches in acoustic, articulatory, and both spaces: Procrustes matching (a physiological approach in articulatory space), vocal tract length normalization (a data-driven approach in acoustic space), and feature space maximum likelihood linear regression (a model-based approach for both spaces), to address the issue of high degree of variation of articulation across different speakers. A preliminary ALS data set was collected and used to evaluate the approaches. Two recognizers, Gaussian mixture model (GMM) - hidden Markov model (HMM) and deep neural network (DNN) - HMM, were used. Experimental results showed adding articulatory data significantly reduced the phoneme error rates (PERs) using any or combined normalization approaches. DNN-HMM outperformed GMM-HMM in all configurations. The best performance (30.7% PER) was obtained by triphone DNN-HMM + acoustic and articulatory data + all three normalization approaches, a 15.3% absolute PER reduction from the baseline using triphone GMM-HMM + acoustic data.

Index Terms: Dysarthric speech recognition, Procrustes matching, vocal track length normalization, fMLLR, hidden Markov models, deep neural network

1. Introduction

Although automatic speech recognition (ASR) technologies have been commercially available for healthy talkers, these technologies did not perform satisfactorily well when directly used for talkers with dysarthria, a motor speech disorder due to neurological or other injury [1]. Dysarthric speech is always with degraded speech intelligibility due to impaired voice and articulation functions [1–3]. For example, Parkinson's disease and amyotrophic lateral sclerosis (ALS) impact the patient's motor functions and therefore impair their speech. Only a few studies have been focused on dysarthric speech recognition [4–6]. Recent studies using mixed data from a variety of neurological diseases indicated articulatory data can improve the speech recognition performance [7, 8]. However, dysarthric speech recognition particularly for ALS has rarely been studied.

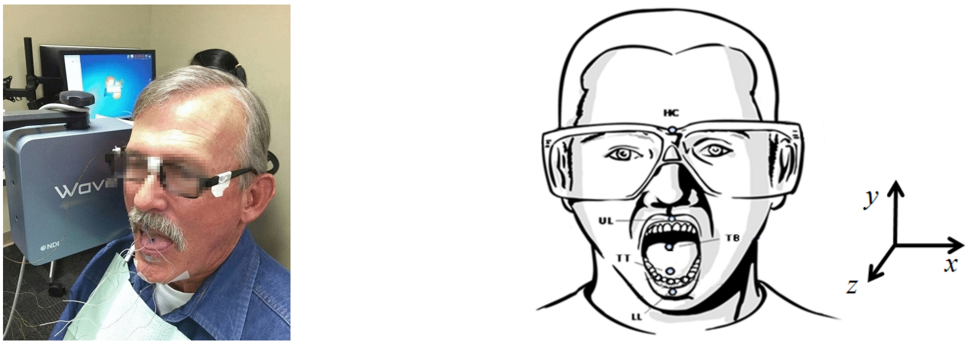
ALS, also known as Lou Gehrig's disease, is the most common motor neuron disease that causes the death of both upper and lower motor neurons [9]. The cause of the disease is unknown for most of the patients and only a small portion (5-10%) of patients is inherited [10]. As the disease progresses, the patient's speech intelligibility declines [11, 12]. Eventually all patients have degraded speech and need an assistive device for communication [13]. Normal speech recognition technology (typically trained on healthy talkers' data) does not work satisfactorily well for the patients. Therefore, ALS patients' ability to use modern speech technology (e.g., smart home environment control driven by speech recognition) is limited. This project, to our best knowledge, is the first one specifically designed to improve speech recognition performance for ALS using articulatory data.

Based on the recent literature on speech recognition with articulatory data (e.g., [7, 14–20]), we hypothesized the followings for dysarthric speech recognition for ALS: 1) adding articulatory data (collected from ALS patients) would improve the speech recognition performance, 2) feature normalization in articulatory, acoustic, and both spaces is critical and necessary for speaker-independent dysarthric speech recognition with articulatory data, and 3) recent state-of-the-art approach, deep neural network (DNN)-hidden Markov model (HMM) would outperform the long-standing approach, Gaussian mixture model (GMM)-HMM.

The high degree of variation in articulatory patterns across speakers has been a barrier for speaker-independent speech recognition with articulatory data. Multiple sources contributed to the inter-talker variation including gender, dialect, individual vocal tract anatomy, and different co-articulation patterns [21]. However, speaker-independent approaches are important for reducing the amount of training data required from each user. Only limited articulatory data samples are often available from individuals with ALS (even with healthy talkers) due to the logistic difficulty of articulatory data collection [22]. For example, in data collection using electromagnetic articulograph (EMA), small sensors have to be attached on the tongue using dental glue [23]. The procedure requires the patient to hold his/her tongue to a position for a while so that the glue can take effect.

To reduce speaker-specific difference, researchers have tried different approaches to normalize the articulatory movements including data-driven approaches (e.g., principal component analysis [7]) or physiological approaches including aligning the tongue position when producing vowels [24–26], consonants [27, 28], and pseudo-words [29] to a reference (e.g., palate [24, 25], or a general tongue shape [27]).

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 47–54,
Dresden, Germany, 11 September, 2015.



(a) *Wave System*

(b) *Sensor Locations. Labels are described in text.*

Figure 1: *Data collection setup.*

Procrustes matching, a bidimensional shape analysis technique [30], has been used to minimize the translational, scaling, and rotational effects of articulatory data across speakers [28, 29, 31]. Recent studies indicated Procrustes matching was effective for speaker-independent silent speech recognition (i.e., recognizing speech from articulatory data only) [18, 19]. Procrustes matching, however, has rarely been used in dysarthric speech recognition with articulatory data.

In addition, we adopted two other representative approaches for across-speaker data normalization. Vocal tract length normalization (VTLN) which has been widely used in acoustic speech recognition [32–36], a data-driven approach in acoustic space, was used to extract normalized acoustic features. The third approach, feature space maximum likelihood linear regression (fMLLR), a model-based adaptation, was used for both acoustic and articulatory data.

In this paper, we investigated the use of 1) articulatory data as additional information source for speech, 2) Procrustes matching, VTLN, and fMLLR as feature normalization approaches individually or combined, 3) two machine learning classifiers, GMM-HMM and DNN-HMM. The effectiveness of these speaker-independent dysarthric speech recognition approaches were evaluated with a preliminary data collected from multiple early diagnosed ALS patients.

2. Data Collection

The dysarthric speech and articulatory data used in this experiment were part of an ongoing project that targets to assess the motor speech decline due to ALS [12, 37].

2.1. Participants and stimuli

Five patients with ALS (3 females and 2 males), American English talkers, participated in the data collection (Table 1). They are all early diagnosed (within half to one year). Severity of these participants with ALS was mild with average speech intelligibility of 94.54% (SD=3.40), with SPK2 not measured. The average age of the patients was 59.80 (SD=7.73). During each session, each subject produced up to 2 or 4 repetitions of 20 unique sentences at their normal speaking rate and loudness. These sentences are used in daily conversations (e.g., *How are you?*) or related to patients (e.g., *This is an emergency*, *I need to see a doctor.*). Some of the sentences were selected from [18, 38].

2.2. Tongue motion tracking device - Wave

The Wave system (NDI Inc., Waterloo, Canada) was used to register the 3-dimensional (x, y, and z; lateral, vertical, and anterior-posterior axes) movements of the tongue and lips during speech production (Figure 1a). Our previous studies [39–41] found four articulators, tongue tip, tongue body back, upper lip, and lower lip, are optimal for this application. Therefore, we used the optimal four sensors for data collection. One sensor was attached on the subject's head and the data were used to calculate the movements of other articulators independent of the head [42]. Wave records tongue movements by establishing a calibrated electromagnetic field that induces electric current into tiny sensor coils that are attached to the surface of the articulators. A similar data collection procedure has been used in [22, 23, 38]. The spatial precision of motion tracking using Wave is approximately 0.5 mm [43]. The sampling rate for recording was 100 Hz.

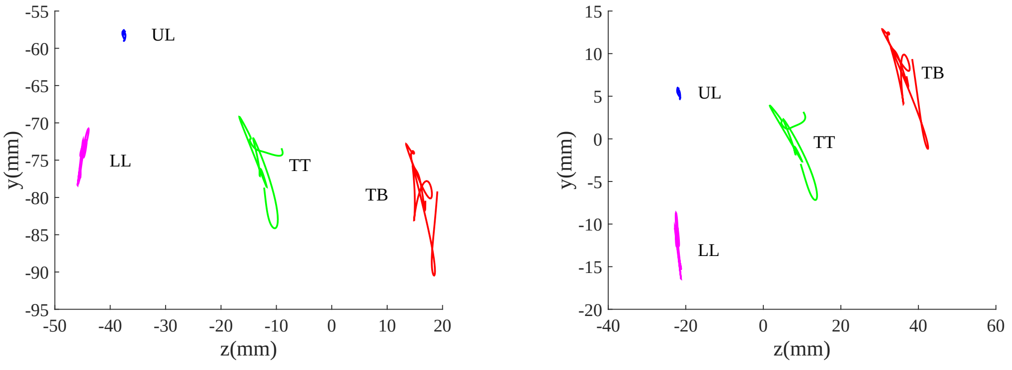
2.3. Procedure

Participants were seated with their head within a calibrated magnetic field (right next to the textbook-sized magnetic field generator). Five sensors were attached to the surface of each articulator using dental glue (PeriAcryl 90, GluStitch) or tape, including one on the head, two on the tongue and two on the lips. A three-minute training session helped the participants to adapt to the wired sensors before the formal data collection.

Figure 1b shows the positions of the five sensors attached to a participant's head, tongue, and lips. HC (Head Center) was on the bridge of the glasses. The movements of HC were used to calculate the head-independent movements of other articulators. TT (Tongue Tip) and TB (Tongue Body Back) were attached at the mid-line of the tongue [22]. TT was about approximately 10 mm from the tongue apex. TB was as far back as possible and about 30 to 40 mm from TT [22]. Lip sensors were attached to the vermilion borders of the upper (UL) and lower (LL) lips at mid-line. Data collected from TT, TB, UL, and LL were used

Table 1: *ALS participants and data size information.*

	Gender	Age	# Phrases	# Frames
SPK1	Female	53	39	5776
SPK2	Female	71	39	5219
SPK3	Male	61	79	9463
SPK4	Female	52	80	13625
SPK5	Male	62	79	9520
Total			316	43603



(a) *Original Data*

(b) *After Procrustes Matching*

Figure 2: *Example of a shape (motion path of four articulators; TT, TB, UL, and LL of SPK5) for producing "Call me back when you can". In this coordinate system, y is vertical and z is anterior-posterior.*

for analysis.

2.4. Data processing

Data processing was applied on the raw sensor position data prior to analysis. First, the head translations and rotations were subtracted from the tongue and lip data to obtain head-independent tongue and lip movement data. The orientation of the derived 3D Cartesian coordinates system is displayed in Figure 1b, in which x is left-right, y is vertical, and z is front-back. Second, a low pass filter (i.e., 20 Hz) was applied for removing noise [22, 23].

In total, 316 sentence samples (for unique twenty phrases) were obtained from the five participants and were used for analysis. It could be expected ALS patients have different lateral movement patterns with healthy subjects (x in Figure 1b) [22], however for this study only y and z coordinates of the tongue and lip sensors were used for analysis.

3. Method

3.1. Procrustes matching: A physiological approach for articulatory data

Procrustes matching (or Procrustes analysis [30]) is a robust statistical bidimensional shape analysis technique, where a shape is represented by a set of ordered landmarks on the surface of an object. Procrustes matching aligns two objects by removing the locational, rotational, and scaling effects [22, 29, 31].

In this project, Procrustes matching was used to match the physiological inter-talker difference (tongue and lip orientation). The downsampled time-series multi-sensor and multi-dimensional articulatory data form articulatory shapes. An example is shown in Figure 2 [18]. This shape contains trajectories of the continuous motion paths of four sensors attached on tongue and lips, TT, TB, UL, and LL. A step-by-step procedure of Procrustes matching between two shapes includes (1) aligning the centroids of the two shapes, (2) scaling the shapes to a unit size, and (3) rotating one shape to match the other [19, 22, 31].

Let S be a set of landmarks as shown below.

$$S = \{(y_i, z_i)\}, \quad i = 1, \ldots, n \tag{1}$$

where (y_i, z_i) represents the i-th data point (spatial coordinates) of a sensor, and n is the total number of data points, where y is vertical and z is front-back. The transformation in Procrustes matching is described using parameters $\{(c_y, c_z), (\beta_y, \beta_z), \theta\}$:

$$\begin{bmatrix} \bar{y}_i \\ \bar{z}_i \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \beta_y \\ \beta_z \end{bmatrix} \begin{bmatrix} y_i - c_y \\ z_i - c_z \end{bmatrix} \tag{2}$$

where (c_y, c_z) are the translation factors (centroids of the two shapes); Scaling factor β is the square root of the sum of the squares of all data points along the dimension; θ is the angle to rotate [30].

Each participant's articulatory shape was transformed into an "normalized shape", which had a centroid at the origin $(0, 0)$ and aligned to the vertical line formed by the average positions (centroids) of the upper and lower lips. Scaling was not used in this experiment, because preliminary tests indicated scaling will cause slightly worse performance in speaker-independent dysarthric speech recognition.

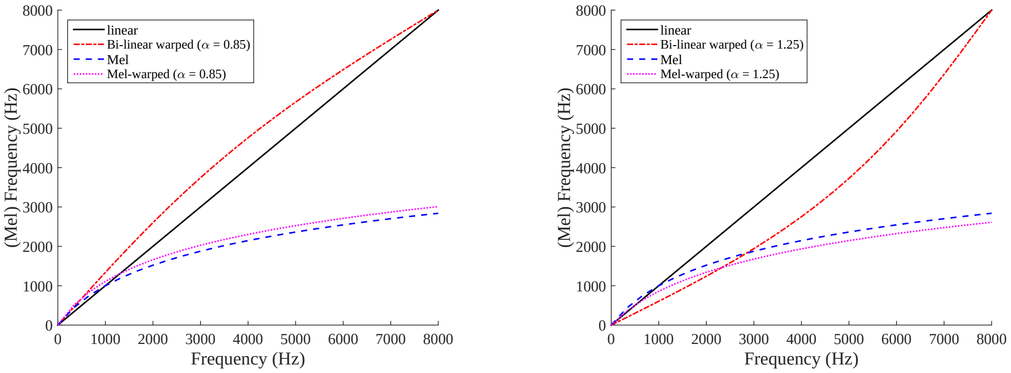
The normalization procedure was done in two steps. First, all articulatory data (e.g., a shape in Figure 2) of each speaker were translated to the centroid (average position of all data points in the shape). This step removed the locational effects between speakers. Second, all shapes of speakers were rotated to make sure the sagittal plane was oriented such that the centroid of lower and upper lip movements defined the vertical axis. This step reduces the variation of rotational effects due to the difference in facial anatomy between speakers. Thus in Eq. 2, (c_y, c_z) are the centroid of shape S; Scaling factor (β_y, β_z) is set to $[1\ 1]'$; θ is the angle of the S to the reference shape in which upper and lower lips form a vertical line. Figure 2 shows an example, original data (Figure 2a) and the shape after Procrustes matching (Figure 2b).

3.2. Vocal tract length normalization: A data-driven approach for acoustic data

Vocal tract length normalization is a representative approach to normalize speaker-dependent characteristics for speech recognition systems [32–36]. This approach is to normalize vocal tract length indirectly from acoustic data, because vocal tract length is highly relevant with pitch and formants [34]. Warping factor α is applied in linear frequency space by Bilinear rule,

$$\hat{F} = F + 2\tan^{-1}\left(\frac{(1-\alpha)\sin(F)}{1-(1-\alpha)\cos(F)}\right) \tag{3}$$

where F is normalized frequency (i.e., divided by sampling frequency, F_s) and α is the warping factor and $F = w/(2\pi F_s)$. Warped Mel-frequency is calculated by applying warping factor



(a) *Warping factor = 0.85 (minumum)* (b) *Warping factor = 1.25 (maximum)*

Figure 3: *Example of (Mel) warped frequency scale (sampling rate: 16 kHz).*

α in Mel-frequency space,

$$M_\alpha(w) = 2595 \log_{10}\left(1 + \frac{w}{\alpha_0 \alpha}\right) \tag{4}$$

where α_0 is 1400π [34] and $w = 2\pi f$ (f: raw frequency). Figure 3 shows an example of (Mel) warped frequency scale between 0.85 and 1.25, the range obtained through empirical studies [34, 44].

In this work, we used linear transformation-based VTLN approach in cepstral space (MFCCs) [35, 36, 44], which was proved equivalent to the above approach [32, 34, 45].

3.3. fMLLR: A model-based approach for both articulatory and acoustic data

fMLLR (also called CMLLR; constrained maximum likelihood linear regression) is one of the representative approaches for across-speaker feature space normalization.

For each speaker, a transformation matrix $\boldsymbol{A}$ and a bias vector $\boldsymbol{b}$ are estimated and used for feature vector transformation:

$$\hat{\boldsymbol{o}}(t) = \boldsymbol{A}\boldsymbol{o}(t) + \boldsymbol{b} \tag{5}$$

where $\boldsymbol{o}(t)$ is the input feature vector at frame t and is transformed to $\hat{\boldsymbol{o}}(t)$. This transformed $\hat{\boldsymbol{o}}(t)$ is used for training GMM-HMM or DNN-HMM and also for decoding. A more detailed explanation of fMLLR can be found in [46].

3.4. Combination of normalization approaches

Besides the individual use of each normalization approach above, we also investigated combinations of these approaches. In this paper, speaker adaptive training (SAT) [46, 47] was conducted using 1) Procrustes matching, VTLN, or fMLLR individually, and 2) combinations with these approaches. We assume the speaker labels for observation are known for training stage. In testing stage, input feature vectors were also transformed using normalization approach(es) as we used in training before they were fed into GMM-HMM or DNN-HMM.

3.5. Recognizer and experimental setup

The long-standing GMM-HMM and recently available DNN-HMM were used as the recognizers [16, 20, 44, 48–50]. In this experiment, window size was 25 ms for acoustic features and frame rate was 10 ms for both acoustic and articulatory features. For each frame, static features plus derivative and acceleration form 39-dimensional mel-frequency cepstral coefficient (MFCC) vectors for acoustic features and 24-dimensional vectors for articulatory features, and these were fed into GMM-HMM or DNN-HMM. HMM is left-to-right 3-state with a monophone or a triphone context model. Maximum likelihood estimation (MLE) training approach (with or without SAT) was used for training GMM-HMM. The input layer of DNN has 216 (24 × 9 frames – 4 previous plus current plus 4 succeeding frames) dimensions for articulatory features and 351 (39 × 9 frames) dimensions for acoustic features. The output layer has 113 dimensions (36 phonemes × 3 states + 1 silence × 5 states) and approximately 200 dimensions (varies for each configuration in triphone model) for monophone and triphone models, respectively. We used 1 to 6 hidden layers and each layer had 512 nodes. The best performance obtained using 1 to 6 layers was

Table 2: *Experimental setup.*

Acoustic Feature	
Feature vector	MFCC (13-dim. vectors) + Δ + $\Delta\Delta$ (39 dim.)
Sampling rate	16 kHz
Windows length	25 ms
Articulatory Feature	
Feature vector	articulatory movement vector (8 dim.) + Δ + $\Delta\Delta$ (24 dim.)
Low pass filtering	20 Hz cutoff 5th order Butterworth
Sampling rate	100 Hz
Concatenated Feature	
Feature vector	MFCC + articulatory movement vector (21 dim.) + Δ + $\Delta\Delta$ (63 dim.)
Common	
Frame rate	10 ms
Mean normalization	Applied
GMM-HMM topology	
Monophone	113 states (36 phones × 3 states, 5 states for silence), total ≈ 1000 mixtures
Triphone	≈ 200 states, total ≈ 1750 mixtures 3-state left to right HMM
Training method	Maximum likelihood estimation (MLE) with and without SAT
DNN-HMM topology	
Input layer dim.	216 (articulatory) 351 (acoustic) 567 (concatenated)
Output layer dim.	113 (monophone) ≈ 200 (triphone)
No. of nodes	512 nodes for each hidden layer
Depth	1 to 6-depth hidden layers
Training method	RBM pre-training, back-propagation
Language model	bi-gram phoneme language model

Table 3: *Angles (in degrees) and centroids (C_y and C_z) in Procrustes matching for each patient.*

	SPK1	SPK2	SPK3	SPK4	SPK5
Angle	34.20°	32.70°	22.11°	22.85°	25.41°
C_y	-62.26	-63.31	-73.29	-71.89	-71.95
C_z	-33.51	-40.89	-26.32	-30.61	-19.60

Note: The degree indicates a counterclockwise rotation. Radians converted from degrees were actually used in the rotation.

Table 4: *Warping factor (α) for each speaker in testing or training stages.*

	CV1	CV2	CV3	CV4	CV5
SPK1	**0.94**	0.95	0.96	0.94	0.99
SPK2	0.93	**0.95**	0.94	0.92	0.98
SPK3	1.01	1.01	**0.99**	0.99	1.04
SPK4	0.95	0.95	0.97	**0.94**	1.00
SPK5	1.05	1.05	1.07	1.05	**1.06**

Note: Diagonal values are for testing and off-diagonal values are for training in each cross-validation (CV). Speakers 1, 2, and 4 are female; speakers 3 and 5 are male.

reported. Table 2 shows the detailed experimental setup. The training and decoding were performed using the Kaldi speech recognition toolkit [44].

Phoneme error rate (PER) was used as the measure of dysarthric speech recognition performance. PER is the summation of substitution, insertion, and deletion errors of phonemes divided by the number of all phonemes.

Leave-one-subject-out cross validation was used in the experiment. In each execution, all samples from one subject were used for testing and the samples from the rest subjects were used for training. The average performance of executions was calculated as the overall performance.

4. Results & Discussion

Table 3 shows detailed parameters (angles and centroids) for Procrustes matching, which varies for different speakers. Table 4 and Figure 4 show the warping factors for each speaker and their histogram. The histogram of ALS patients follows general trend of warping factor distribution for females (typically < 1.0) and males (typically > 1.0).

Figures 5, 6, 7, and 8 give the PERs of speaker-independent dysarthric (due to ALS) speech recognition results using different context models and recognizers, respectively: (1) monophone GMM-HMM, (2) triphone GMM-HMM, (3) monophone DNN-HMM, and (4) triphone DNN-HMM with individual or combinations of VTLN, Procrustes matching, and fMLLR. These results suggest that VTLN, Procrustes matching, and fMLLR were all effective for speaker-independent dysarthric speech recognition from acoustic data, articulatory data, or combined. When comparing the three normalization approaches individually (if applies), no approach was universally better than others in all experimental configurations. A better performance was always obtained when the normalization approaches were combined. Baseline results were obtained without using any normalization approach.

Adding articulatory data to acoustic data always showed performance improvement in all configurations (monophone/triphone or GMM-HMM/DNN-HMM), which is consistent with the literature [7]. The overall best performance was obtained when the three normalization approaches, VTLN (acoustic space), Procrustes matching (articulatory space), and fMLLR (both acoustic and articulatory space), were used together with triphone DNN-HMM model (30.7%).

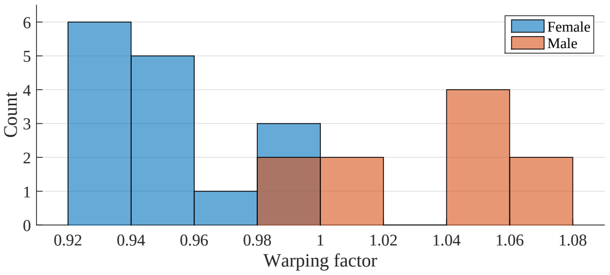


Figure 4: *Histogram of warping factors (step size = 0.02).*

Surprisingly, speaker-independent silent speech recognition (using articulatory data only) with DNN-HMM obtained even better results than the recognition results from acoustic (MFCC) features (see left half of Figures 7 and 8). This finding shows the potential of articulatory data when the patient's speech is significantly impaired as the disease progresses. However, since the data set is small, a further study with a larger data set is required to verify this finding.

Moreover, DNN-HMM outperformed GMM-HMM in all configurations (monophone/triphone, VTLN/Procrustes matching/fMLLR). This finding is consistent with the acoustic [20,51] and silent speech recognition literature [17, 19].

In the current approach, fMLLR was not separately applied to acoustic and articulatory data (i.e., full transformation matrix), because the two types of data are concatenated before applying fMLLR. Due to the different nature of acoustic (in frequency domain) and articulatory data (in spatial domain), in the future, we consider to make $\boldsymbol{A}$ in Eq. 5 a block-diagonal transformation matrix. The block-diagonal matrix will separate the processing for acoustic and articulatory data.

Limitations. Although the experimental results were encouraging, the data set used in the experiment contained only a small number of unique phrases collected from a small number of ALS patients. Further studies with a larger vocabulary from more ALS patients are necessary to explore the limits of the current approaches.

5. Conclusions & Future Work

This paper investigated speaker-independent dysarthric speech recognition using the data from patients with ALS and also with three across-speaker normalization approaches: a physiological approach, Procrustes matching, a data-driven approach, VTLN, and a model-based approach, fMLLR. GMM-HMM and DNN-HMM were used as the machine learning classifiers. Experimental results showed the effectiveness of feature normalization approaches. The best performance was obtained when the three approaches were used together with triphone DNN-HMM.

Future work includes test of the normalization approaches using a larger data set collected from more ALS subjects (e.g, by combining our data set with the ALS data in TORGO [8]).

6. Acknowledgments

This work was supported by the National Institutes of Health through grants R01 DC013547 and R03 DC013990. We would like to thank Dr. Jordan R. Green, Dr. Thomas F. Campbell, Dr. Anusha Thomas, Jana Mueller, Victoria Juarez, Saara Raja, Soujanya Koduri, Kumail Haider, the volunteering partic-

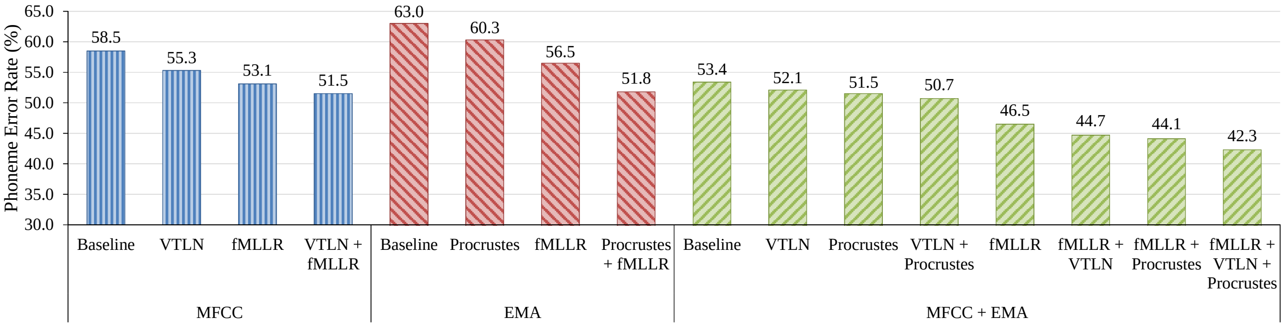


Figure 5: *Phoneme Error Rates (PERs; %) of speaker-independent recognition using* ***monophone GMM-HMM*** *with fMLLR, VTLN, and/or Procrustes matching.*

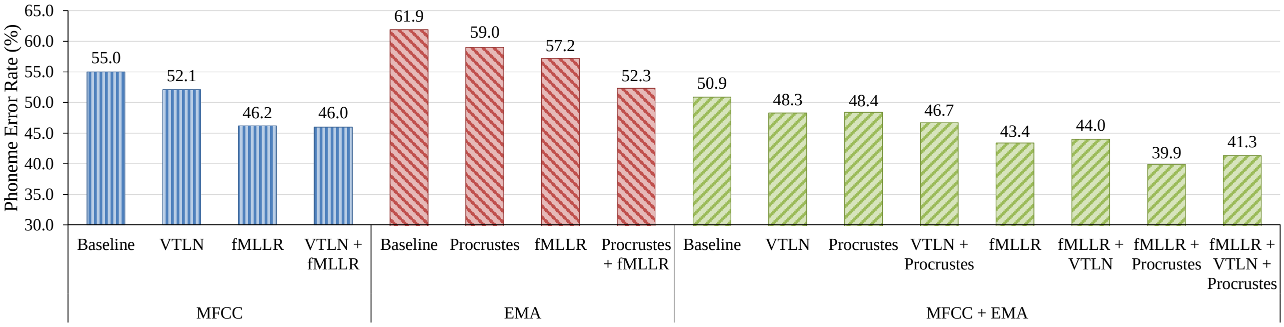


Figure 6: *Phoneme Error Rates (PERs; %) of speaker-independent recognition using* ***triphone GMM-HMM*** *with fMLLR, VTLN, and/or Procrustes matching.*

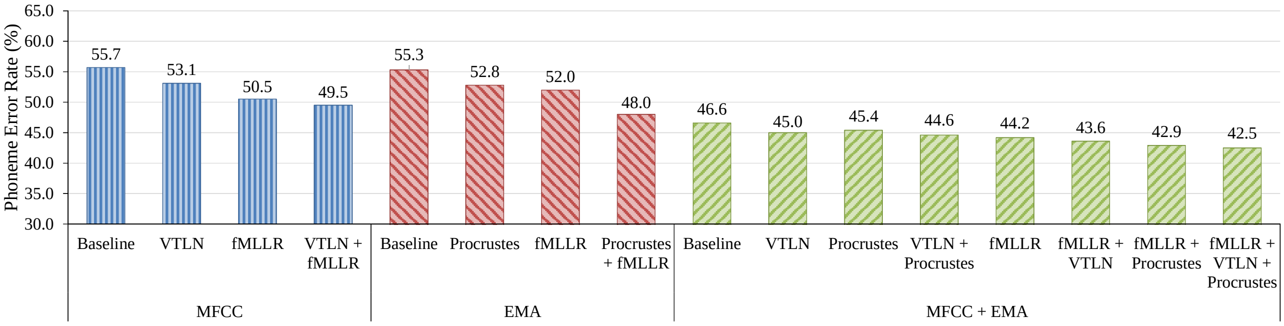


Figure 7: *Phoneme Error Rates (PERs; %) of speaker-independent recognition using* ***monophone DNN-HMM*** *with fMLLR, VTLN, and/or Procrustes matching.*

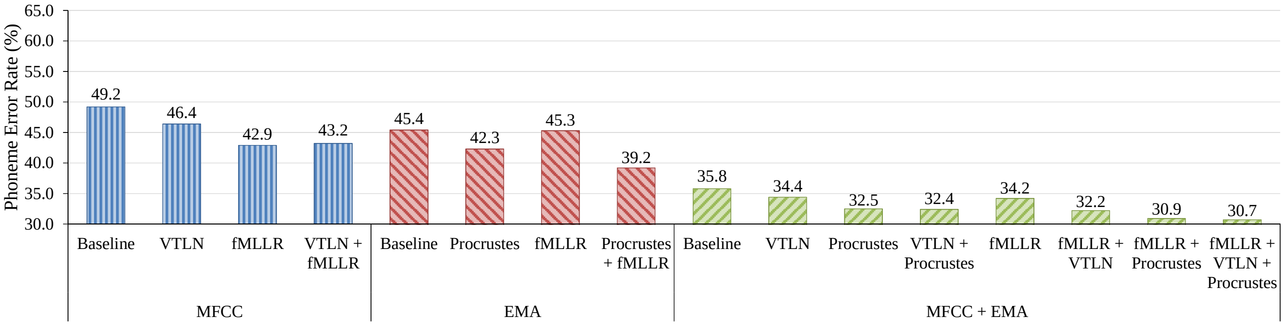


Figure 8: *Phoneme Error Rates (PERs; %) of speaker-independent recognition using* ***triphone DNN-HMM*** *with fMLLR, VTLN, and/or Procrustes matching.*

ipants, and the Communication Technology Center, University of Texas at Dallas.

7. References

[1] J. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Mosby; 3rd edition, 2005.

[2] E. Hanson, K. M. Yorkston, and D. Britton, "Dysarthria in amyotrophic lateral sclerosis: A systematic review of characteristics, speech treatment and AAC options," *Journal of Medical Speech - Language Pathology*, vol. 19, no. 3, pp. 12–30, 2011.

[3] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel contrast and speech intelligibility in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4, pp. 187–194, 2011.

[4] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech and Language*, vol. 27, no. 6, pp. 1147–1162, 2013.

[5] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP J. Adv. Signal Process*, vol. 2009, pp. 2:1–2:14, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/308340

[6] K. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4924–4927.

[7] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2011.

[8] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria." *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[9] M. Kiernan, S. Vucic, B. Cheah, M. Turner, A. Eisen, O. Hardiman, J. Burrell, and M. Zoing, "Amyotrophic lateral sclerosis," *Lancet*, vol. 377, no. 9769, pp. 942–955, 2011.

[10] J. Sreedharan, I. P. Blair, V. B. Tripathi, X. Hu, C. Vance, B. Rogelj, S. Ackerley, J. C. Durnall, K. L. Williams, E. Buratti, F. Baralle, J. de Belleroche, J. D. Mitchell, P. N. Leigh, A. Al-Chalabi, C. C. Miller, G. Nicholson, and C. E. Shaw, "TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis," *Science*, vol. 319, no. 5870, pp. 1668–1672, 2008.

[11] B. R. Brooks, R. Miller, M. Swash, and T. Munsat, "El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotroph Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 1, no. 5, pp. 293–299, 2000.

[12] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, pp. 494–500, 2013.

[13] D. Beukelman, S. Fager, and A. Nordness, "Communication support for people with ALS," *Neurology Research International*, no. 714693, p. 6 pages, 2011.

[14] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[15] J. Wang, "Silent speech recognition from articulatory motion," Ph.D. dissertation, The University of Nebraska-Lincoln, 2011.

[16] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Cross-corpus and cross-linguistic evaluation of a speaker-dependent DNN-HMM ASR system using EMA data," in *Proc. of Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, 2013.

[17] S. Hahm and J. Wang, "Silent speech recognition from articulatory movements using deep neural network," in *Proc. of the 18th Intl. Congress of Phonetic Sciences*, 2015.

[18] J. Wang, A. Samal, and J. Green, "Across-speaker articulatory normalization for speaker-independent silent speech recognition," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 1179–1183.

[19] J. Wang and S. Hahm, "Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training," in *Proc. of INTERSPEECH*, 2015.

[20] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping." in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 1297–1301.

[21] R. Kent, S. Adams, and G. Tuner, "Models of speech production," in *Principles of Experimental Phonetics, (Lass, N.J., ed.)*, 1996, pp. 3–45, Mosby.

[22] J. Wang, J. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[23] J. Green, J. Wang, and D. L. Wilson, "Smash: A tool for articulatory data processing and analysis," in *Proc. of INTERSPEECH*, 2013, pp. 1331–1335.

[24] K. Johnson, P. Ladefoged, and M. Lindau, "Individual differences in vowel production," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 701–714, 1993.

[25] M. Hashi, J. R. Westbury, and K. Honda, "Vowel posture normalization," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2426–2437, 1998.

[26] A. P. Simpson, "Gender-specific differences in the articulatory and acoustic realization of interword vowel sequences in american english," in *5th Seminar on Speech Production: Models and Data. Kloster Seeon*, 2000, pp. 209–212.

[27] J. R. Westbury, M. Hashi, and M. J Lindstrom, "Differences among speakers in lingual articulation for American English /ɹ/," *Speech Communication*, vol. 26, no. 3, pp. 203–226, 1998.

[28] S. Li and L. Wang, "Cross linguistic comparison of Mandarin and English EMA articulatory data," in *Proc. of INTERSPEECH*, 2012, pp. 903–906.

[29] D. Felps, S. Aryal, and R. Gutierrez-Osuna, "Normalization of articulatory data through procrustes transformations and analysis-by-synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3027–3031.

[30] I. L. Dryden and K. V. Mardia, *Statistical shape analysis*. John Wiley & Sons New York, 1998, vol. 4.

[31] J. Wang, J. R. Green, A. Samal, and D. B. Marx, "Quantifying articulatory distinctiveness of vowels," in *Proc. of INTERSPEECH*, 2011, pp. 277–280.

[32] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. of ICASSP*, vol. 1, 1996, pp. 346–348.

[33] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Proc. of ICASSP*, vol. 2, 1997, pp. 1039–1042.

[34] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," *CMU-CS-97-148, Carnegie Mellon University, Pittsburgh, PA*, 1997.

[35] T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," in *Proc. of Eurospeech*, 2001, pp. 1649–1652.

[36] D. Kim, S. Umesh, M. Gales, T. Hain, and P. Woodland, "Using VTLN for broadcast news transcription," in *Proc. of INTERSPEECH*, 2004, pp. 1953–1956.

[37] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, "Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behavioral Neurology*, no. 183027, pp. 1–11, 2015.

[38] J. Wang, A. Samal, J. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 4985–4988.

[39] J. Wang, J. Green, and A. Samal, “Individual articulator’s contribution to phoneme production,” in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7785–7789.

[40] J. Wang, S. Hahm, and T. Mau, “Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition,” in *ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015.

[41] J. Wang, A. Samal, P. Rong, and J. R. Green, “An optimal set of flesh points on tongue and lips for speech movement classification,” *Journal of Speech, Language, and Hearing Research*, In press.

[42] J. R. Green and Y.-T. Wang, “Tongue-surface movement patterns during speech and swallowing,” *The Journal of the Acoustical Society of America*, vol. 113, no. 5, pp. 2820–2833, 2003.

[43] J. Berry, “Accuracy of the NDI wave speech research system.” *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1295–301, 2011.

[44] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and V. K., “The Kaldi speech recognition toolkit,” in *Proc. of ASRU*, Waikoloa, USA, 2011, pp. 1–4.

[45] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[46] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer speech and language*, vol. 12, no. 2, pp. 75–98, 1998.

[47] T. Anastasakos, J. McDonough, and J. Makhoul, “Speaker adaptive training: A maximum likelihood approach to speaker normalization,” in *Proc. of ICASSP*, vol. 2, 1997, pp. 1043–1046.

[48] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[49] G. Hinton, “A practical guide to training restricted Boltzmann machines,” *Momentum*, vol. 9, no. 1, p. 926, 2010.

[50] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, “Recent advances in deep learning for speech research at microsoft,” in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 8604–8608.

[51] A.-R. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

Vowel Enhancement in Early Stage Spanish Esophageal Speech Using Natural Glottal Flow Pulse and Vocal Tract Frequency Warping

Rizwan Ishaq[1], *Dhananjaya Gowda*[2], *Paavo Alku*[2], *Begoña García Zapirain*[1]

[1]Deustotech-LIFE, University of Deusto, Bilbao, Spain
[2]Aalto University, Dept. of Signal Processing and Acoustics, Finland

rizwanishaq@deusto.es, dhananjaya.gowda@aalto.fi, paavo.alku@aalto.fi, mbgarciazapi@deusto.es

Abstract

This paper presents an enhancement system for early stage Spanish Esophageal Speech (ES) vowels. The system decomposes the input ES into neoglottal waveform and vocal tract filter components using Iterative Adaptive Inverse Filtering (IAIF). The neoglottal waveform is further decomposed into fundamental frequency F_0, Harmonic to Noise Ratio (HNR), and neoglottal source spectrum. The enhanced neoglottal source signal is constructed using a natural glottal flow pulse computed from real speech. The F_0 and HNR are replaced with natural speech F_0 and HNR. The vocal tract formant frequencies (spectral peaks) and bandwidths are smoothed, the formants are shifted downward using second order frequency warping polynomial and the bandwidth is increased to make it close to the natural speech. The system is evaluated using subjective listening tests on the Spanish ES vowels /a/, /e/, /i/, /o/, /u/. The Mean Opinion Score (MOS) shows significant improvement in the overall quality (naturalness and intelligibility) of the vowels.

Index Terms: speech enhancement, glottal flow, analysis synthesis vocal tract, spectral sharpening, warping

1. Introduction

The removal of the larynx after a Total Laryngectomy (TL), changes the speech production mechanism. The trachea which connects the larynx and lungs for air source is now connected to a stoma (hole on neck) for breathing. The vocal folds which resided in larynx are no more available. After TL, there is no voicing and air source for speech production. Therefore alternative voicing and air source are needed for speech restoration. Three methods are available for this purpose, i) Esophageal Speech (ES), ii) Tracheo-Esophageal Speech (TES), and iii) Electrolarynx (EL). ES and TES both use a common voicing source, the Phyarngo-Esophageal (PE) segment, but with a different air source, while EL uses external devices for voicing source with no air source. The ES is preferred over other methods, because it does not require surgery (TES) or external devices (EL). ES involves, however, a low pressure air source, and an irregular PE segment vibration which results in low quality and low intelligible speech. Compared to the production of normal speech according to the source-filter model [1], the voicing source in ES is severely altered and does not have any fundamental frequency or harmonic components. The vocal tract filter is also shortened in ES. The ES can be enhanced by transforming the source and filter components to those of normal speech using signal processing algorithms.

In previous studies ES is typically decomposed into its source and filter components using Linear Predication (LP) based analysis-synthesis techniques. Based on this assumption the authors in [2, 3] replaced the voicing source with the Liljencrants- Fant (LF) voicing source, and reported significant enhancements. Fundamental frequency smoothing and correction with the synthetic LF source model were used for quality enhancement also in [4]. ES enhancement based on formant synthesis has also shown significant improvement in intelligibility [5, 6]. In [7] the source and filter components were modified by replacing the source with the LF model and increasing the bandwidth of filter formants for better quality speech. Statistical conversion from ES to normal speech has also improved intelligibility, but requires more ES data [8]. Some other not so common approaches are based on Kalman filtering [9, 10, 11, 12], and modulation filtering enhancement [13, 14].

Almost all methods available in the literature assume that the fundamental frequency of ES can be estimated accurately. The voicing source signal is then modified with the synthetic LF model voicing source. The vocal tract formants are typically considered to be the same as in normal speech signals. In reality, however, the fundamental frequency of ES is highly irregular and the voicing source resembles whispered speech. Moreover, formants center frequencies are affected by the shortening of vocal tract length due to surgery. In order to deal with these deficiencies, this paper proposes an ES enhancement method based on the GlottHMM single pulse synthesis [15, 16, 17]. The system decomposes ES into neoglottal waveform and vocal tract filter components using Iterative Adaptive Inverse Filtering (IAIF) [18]. Natural glottal pulse extracted from real speech is used to construct the glottal waveform by borrowing F_0 curve and HNR from normal speech. The vocal tract filter is also modified by smoothing the spectral peaks and their bandwidths. The spectral peaks of the vocal tract filter are also moved to lower frequencies in order to compensate the rising of formant in ES. The formant bandwidths are also increased for better quality speech. The system is validated with Spanish Esophageal Vowels subjectively using the Mean Opinion Score (MOS). The paper in next section describes the system in detail. The subsequent sections contain results, discussion and finally conclusions.

2. System Description

The proposed system, shown in Figure 1, is divided into three main components, i) analysis, ii) transformation, and iii) synthesis. The analysis part decomposes the voiced speech frame into its source and filter components. The transformation provides the modified source and filter components. Finally the modified components are combined in the synthesis part to generate enhanced ES.

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 55–59,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

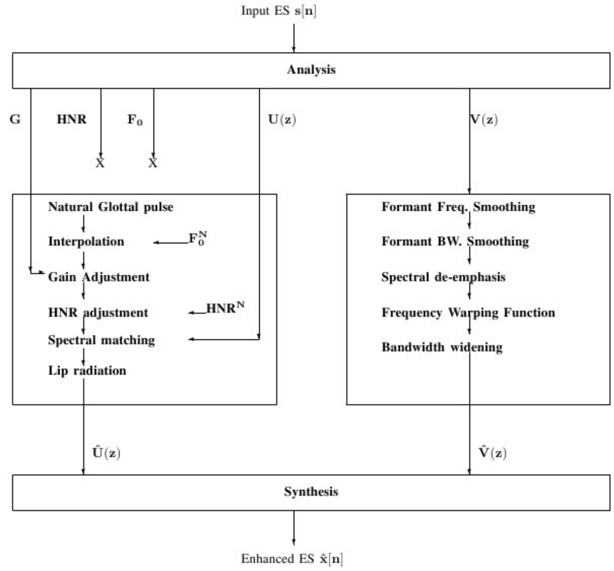


Figure 1: *Proposed enhancement system.*

2.1. GlottHMM based analysis

The goal of the analysis part of the system is to decompose the ES signal into a neoglottal source signal and a vocal tract spectrum. The input speech signal $s[n]$ is first passed through high-pass filter $h_{hp}[n]$ with a cutoff frequency of 70 Hz.

$$s_h[n] = s[n] * h_{hp}[n] \tag{1}$$

where $s_h[n]$ and $*$ are the highpass filtered speech signal and a convolution operator, respectively. The highpass filtered signal $s_h[n]$ is then windowed using a rectangular window of size 45-ms, with 5-ms frame shift.

$$x[n] = s_h[n]w[n] \tag{2}$$

where $w[n]$ is the rectangular window. Firstly the log energy G of frame is extracted using,

$$G = log(\sum_{n=0}^{N-1} x^2[n]) \tag{3}$$

where N is the number of samples in the frame. Glottal Inverse Filtering (GIF) is then used to separate the frame into a neoglottal source signal and a vocal tract spectrum. The automatic inverse filtering, IAIF is used [18]. IAIF estimates vocal tract and lip radiation using all-pole modeling and then iteratively cancel these components. In simplified form, the neoglottal source signal:

$$U(z) = \frac{X(z)}{V(z)R(z)} \tag{4}$$

where $U(z)$, $X(z)$, $V(z)$ and $R(z)$ are the z-transforms of neoglottal source signal $u[n]$, speech signal $x[n]$, vocal tract impulse response $v[n]$, and lip radiation response $r[n]$ respectively. The estimated neoglottal source signal $u[n]$ is parametrized into fundamental frequency F_0, Harmonic to Noise Ratio (HNR) and neoglottal source spectrum $U(z)$. The autocorrelation of the neoglottal source signal $u[n]$ is used for F_0 estimation. The HNR is estimated using the upper and lower smoothed spectral envelopes ratio to determine the voicing degree in the neoglottal voicing source signal $u[n]$ for five frequency bands [15]. In short the analysis part of the system provides for each frame the following, i) Frame energy G, ii) vocal tract spectrum $V(z)$ (LP order 30), iii) F_0, iv) HNR and v) neoglottal source spectrum $U(z)$ (LP order 10).

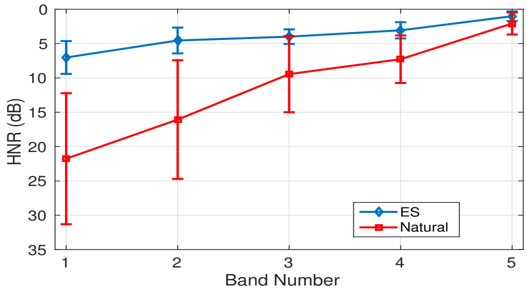


Figure 2: *HNR of ES and natural speech.*

2.2. ES to normal speech transformation

The parameters obtained from the analysis are transformed into natural speech parameters. The neoglottal signal and vocal tract are modified independently.

2.2.1. Neoglottal source signal enhancement

The neoglottal source signal $u[n]$ is the most effected speech component in ES. Therefore the parameters of this signal are replaced with any arbitrary natural speech signal for a better glottal source signal. The natural glottal pulse which is extracted from normal speech is first interpolated using the cubic spline interpolation by replacing the frame original F_0 with natural speech F_0^N . The interpolated glottal pulse voicing source is then multiplied with the smooth gain G and the natural speech HNR is then used to add noise in the frequency domain for naturalness according to the following steps:

- Taking FFT of the neoglottal waveform,
- Adding random components (white Gaussian noise) to real and imaginary part of FFT according to HNR,
- Taking IFFT of noise added neoglottal waveform

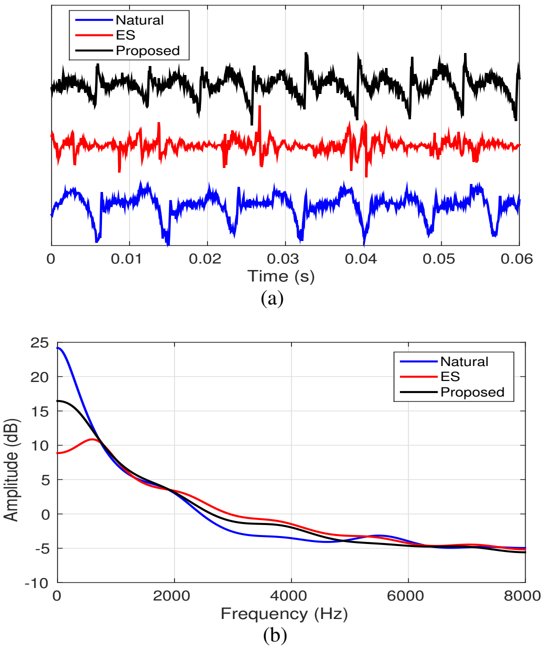
$$U_{syn}(z) = 10^G G(z) + Q(z) \tag{5}$$

where $U_{syn}(z)$ is the synthetic glottal source, $G(z)$ is the natural glottal pulses source, and $Q(z)$ is HNR based noise component. Figure 2 shows the mean value of HNR for all voiced frames along with standard deviation. The figure indicates that HNR of ES is greatly different from that of normal speech. Therefore, it is justified to replace the HNR of ES with the HNR of normal speech in the vowel enhancement system. In order to adjust the spectrum of neoglottal waveform to the spectrum of the target waveform, the former is filtered with following IIR filter:

$$H_m(z) = \frac{U(z)}{U_{syn}(z)} \tag{6}$$

where $U(z)$ and $U_{syn}(z)$ are the LP spectra of the original and synthetic neoglottal waveform, respectively. The lip radiation is applied to the spectrally matched neoglottal waveform $\hat{u}[n]$:

$$\hat{u}[n] = \hat{u}[n] - \alpha\hat{u}[n-1], \quad 0.96 < \alpha < 1 \tag{7}$$



(a)

(b)

Figure 3: *Glottal excitations (computed from the vowel /a/) in the time domain (a) and in the frequency domain (b).*

where $\hat{u}[n](\hat{U}(z))$ and $\alpha(0.98)$ are the modified neoglottal waveform and lip radiation constant, respectively.

Figure 3(a) shows time-domain examples of glottal excitations of natural speech and ES together with a waveform computed with the proposed enhancement system. It can be seen that the proposed system is capable of producing a glottal excitation that is highly similar to that of natural speech. As shown in Figure 3(b), the spectral slope of the excitation waveform generated by the proposed method is also close to that of natural speech, especially at low frequencies, but the generated spectrum also retains the spectral slope of ES at higher frequencies.

2.2.2. *Vocal tract modification by nonlinear frequency warping*

The vocal tract spectrum of ES has the following characteristics, i) higher frequencies are emphasized more compared to lower frequencies, ii) spectral resonances (formants) are moved to higher frequencies, and iii) resonance bandwidths are reduced in comparison to normal speech vowels. To cope with the higher frequency emphasis, a de-emphasis filter is applied to the vocal tract spectrum. The resulting vocal tract transfer function is then expressed as:

$$H_{enh}(z) = \frac{1+\alpha z^{-1}}{1+\sum_{p=1}^{P} a_p z^{-p}}, \quad 0.95 < \alpha < 1 \tag{8}$$

where P is the order of the all-pole vocal tract filter and α is the de-emphsis constant.

Because formants of ES are moved upward in frequency, a procedure is needed to adjust them to coincide more closely with the formant values of normal speech. For such a procedure, we used a second order Frequency Warping Function (FWF) $\zeta(f)$ defined as:

$$\zeta(f) = \alpha_1 f^2 + \alpha_2 f + c \tag{9}$$

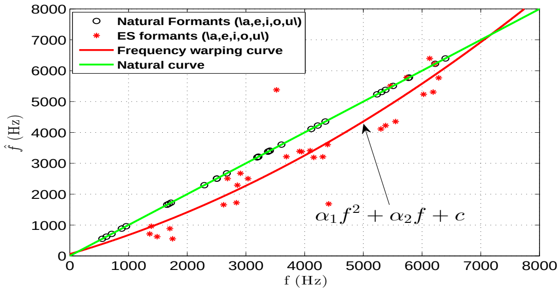


Figure 4: *Frequency Warping Function (FWF) curve.*

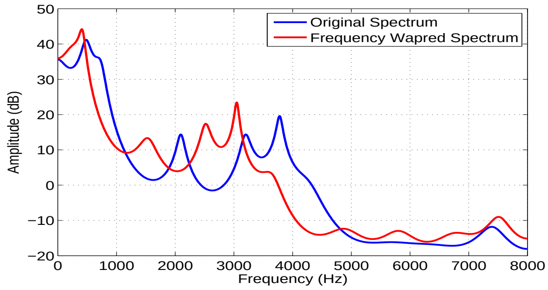


Figure 5: *Frequency warped spectra.*

where $\alpha_1 = 6.079 \times 10^{-5}$, $\alpha_2 = 0.5553$, and $c = 60.280$.

$$\hat{f} = \beta\zeta(f), \quad \beta = 1, f = 0 \rightarrow \frac{f_s}{2} \tag{10}$$

where $\hat{f}$ and f, are warped and original frequencies, and β is a constant. Figure 4 demonstrates FWF using first four formants of vowels (/a/, /e/, /i/, /o/, /u/) extracted from normal speech (x-axis) and ES (y-axis). The obtained frequency warping, applicable for a general formant mapping between normal speech and ES, is shown in Figure 5. In order to expand the formant bandwidths, exponential windowing is used for the vocal tract filter coefficients as follows [19]:

$$H_s(z) = \frac{1+\sum_{p=1}^{P} \gamma^p a_p z^{-p}}{1+\sum_{p=1}^{P} \eta^p a_p z^{-p}}, \quad 0.90 < \gamma, \eta < 1 \tag{11}$$

where γ and η are constants controlling the spectral bandwidth.

If $\gamma > \eta$ bandwidth of formants increase, otherwise it decreases (i.e. formants are sharpened). For the purpose of the present study, $\eta(0.97)$ is always smaller than $\gamma(0.99)$ in order to increase formant bandwidths.

2.3. Synthesis of enhanced speech

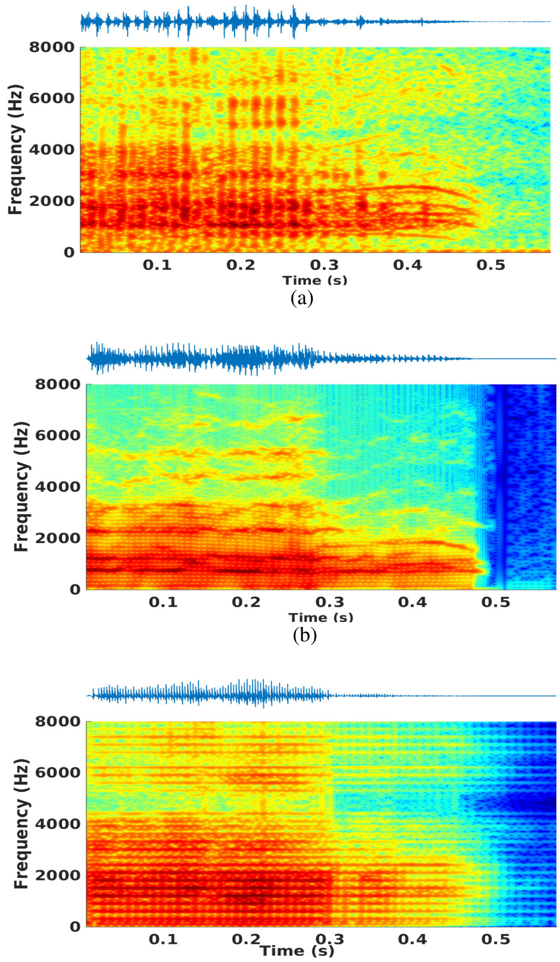
The synthesis part involves convolving the modified neoglottal waveform and the impulse response of the vocal tract filter yielding the enhanced version of ES $\hat{x}[n]$;

$$\hat{x}[n] = \hat{v}[n] * \hat{u}[n] \tag{12}$$

where $\hat{u}[n]$ and $\hat{v}[n]$ are the modified neoglottal waveform and vocal tract impulse response, respectively.

3. System Evaluation

The system was evaluated with ES vowels of Spanish (/a/, /e/, /i/, /o/, /u/) recorded in speech rehabilitation center. The data



(a)

(b)

(c)

Figure 6: *Spectrograms of the vowel /a/ for different processing types: unprocessed (a), processed with the proposed system (b), processed with the reference system (c) [7]*

was collected from five early stage male ES talkers by asking them to utter each vowel four times. Due to lack of female patients in the rehabilitation center, only male speakers were involved in the study. The speech sounds were sampled with 44.1 kHz from which the data was down-sampled to 16 kHz for computational efficiency.

The system performance is visually demonstrated with spectrograms in Figure 6. In this figure, and also later in Figures 7 and 8, the proposed system is compared with a reference system based on using the LF source and formant modification with a bandwidth extension system [7]. It can be seen from Figure 6 that the spectrogram computed from the enhanced vowels by the proposed system shows a clearer formant and harmonics structure in comparison to ES and the reference system.

3.1. Subjective listening evaluation

Two subjective listening tests were conducted. The first one was a quality evaluation based on the Mean Opinion Score (MOS) which is a widely used perceptual quality test of speech based on a scale from 1 (worst) to 5 (best). In this test, the listeners heard original ES vowels and the corresponding enhanced ones, processed by both the proposed and the reference method, in a random order and they were asked to grade the quality of the sounds on the MOS scale. The second listening test was a preference test where the listeners heard vowels corresponding to the same three processing types and they were asked to select which one they prefer to listen. A total of 10 listeners participated in the listening tests.

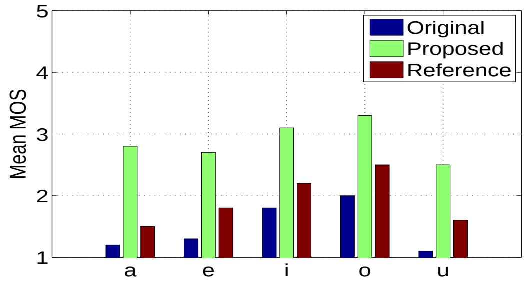


Figure 7: *Results of the MOS test for all the vowels.*

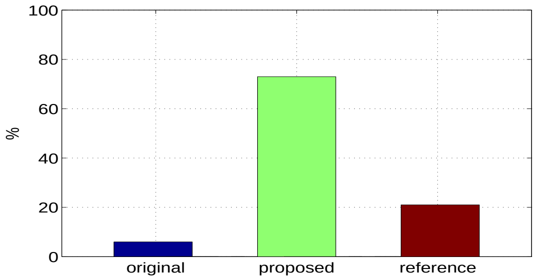


Figure 8: *Results of the preference test.*

Figure 7 shows the results of the MOS test. The data indicates that the proposed system has a mean MOS higher than 2.5 for all the vowels, which can be considered a good quality score for ES samples. Figure 8 shows the data of the preference tests by combining all the vowels. Also these data indicate that the proposed method has succeeded in enhancing the quality of the ES vowels.

4. Conclusion

An enhancement system for ES vowels was proposed based on using a natural glottal pulse combined with second order polynomial Frequency Warping Function. A preliminary evaluation of the system was carried out on early stage Spanish ES vowels by comparing the system performance with a known reference method. Results obtained with a MOS evaluation show clear improvements in speech quality both in comparison to the original ES vowels and to sounds enhanced with the reference method. The good performance was corroborated with a preference test indicating that in the vast majority of the cases, listeners preferred to listen to the sounds enhanced by the proposed method. Future work is needed to study the system together with advanced stage ES speakers.

5. Acknowledgements

Special thanks to all my colleagues at Aalto University for their valuable support and time.

6. References

[1] G. Fant, "Acoustic theory of speech production." Mouton, The Hauge, 1960.

[2] Q. Yingyong, W. Bernd, and B. Ning, "Enhancement of female esophageal and tracheoesophageal speech," *Acoustical Society of America*, vol. 98(5, Pt1), pp. 2461–2465, 1995.

[3] Y. Qi, "Replacing tracheoesophageal voicing source using lpc synthesis," *Acoustical Society of America*, vol. 5, pp. 1228–1235, 1990.

[4] R. Sirichokswad, P. Boonpramuk, N. Kasemkosin, P. Chanyagorn, W. Charoensuk, and H. H. Szu, "Improvement of esophageal speech using lpc and lf model," *Internation Conf. on Biomedical and Pharamaceutical Engineering 2006*, pp. 405–408, 2006.

[5] M. Kenji, H. Noriyo, K. Noriko, and H. Hajime, "Enhancement of esophageal speech using formant synthesis," *Acoustic. Sci. and Tech.*, pp. 69–76, 2002.

[6] M. Kenji and H. Noriyo, "Enhancement of esophageal speech using formant synthesis," *Acoustics, Speech and Signal Processing, International conf.*, pp. 81–85, 1999.

[7] R. H. Ali and S. B. Jebara, "Esophageal speech enhancement using excitation source synthesis and formant structure modification," *SITIS*, pp. 615–624, 2006.

[8] K. Doi, H.and Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Statistical approach to enhancing esophageal speech based on gaussian mixture models," *Acoustics Speech and Signal Processing(ICASSP), 2010 IEEE International Conference*, pp. 4250–4253, 2010.

[9] O. Ibon, B. Garcia, and Z. M. Amaia, "New approach for oesophageal speech enhancement," *10th International conference, ISSPA*, vol. 5, pp. 225–228, 2010.

[10] B. Garcia and A. Mendez, "Oesophageal speech enhancement using poles stablization and kalman filtering," *ICASSP*, pp. 1597–1600, 2008.

[11] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Oesophageal voice acoustic parameterization by means of optimum shimmer calculation," *WSEAS Trasactions on Systems*, pp. 489–499, 2008.

[12] R. Ishaq and B. G. Zapirain, "Optimal subband kalman filter for normal and oesophageal speech enhancement," *Bio-Medical Materials and Engineering*, vol. 24, pp. 3569–3578, 2014.

[13] R. Ishaq, B. G. Zapirain, M. Shahid, and B. Lovstrom, "Subband modulator kalman filtering for signla channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[14] R. Ishaq and B. G. Zapirain, "Adaptive gain equalizer for improvement of esophageal speech," in *IEEE International Symposium on Signal Processing and Information Technology*, 2012.

[15] A. Suni, T. Raitio, , M. Vainio, and P. Alku, "The glottalHMM entery for blizzard challenge 2011: Utilizing source unit selection in hmm-based speech synthesis for improved excitation generation," in *in Blizzard Challenge 2011, Workshop, Florence, Italy*, 2011.

[16] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 153–165, 2011.

[17] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for hmm-based speech synthesis," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2011.

[18] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," in *Speech communication*, vol. 11, no. 2, 1992, pp. 109–118.

[19] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, pp. 59–71, 1995.

Automatic dysfluency detection in dysarthric speech using deep belief networks

Stacey Oue[1], *Ricard Marxer*[2], *Frank Rudzicz*[1,3]

[1]Department of Computer Science, University of Toronto;
[2]Department of Computer Science, University of Sheffield;
[3]Toronto Rehabilitation Institute-UHN

stacey.oue@mail.utoronto.ca, r.marxer@sheffield.ac.uk, frank@cs.toronto.edu

Abstract

Dysarthria is a speech disorder caused by difficulties in controlling muscles, such as the tongue and lips, that are needed to produce speech. These differences in motor skills cause speech to be slurred, mumbled, and spoken relatively slowly, and can also increase the likelihood of dysfluency. This includes non-speech sounds, and 'stuttering', defined here as a disruption in the fluency of speech manifested by prolongations, stop-gaps, and repetitions. This paper investigates different types of input features used by deep neural networks (DNNs) to automatically detect repetition stuttering and non-speech dysfluencies within dysarthric speech. The experiments test the effects of dimensionality within Mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs), and explore the detection capabilities in dyarthric versus non-dysarthric speech. The results obtained using MFCC and LPCC features produced similar recognition accuracies; repetition stuttering in dysarthric speech was identified correctly at approximately 86% and 84% for non-dysarthric speech. Non-speech sounds were recognized with approximately 75% accuracy in dysarthric speakers.

Index Terms: Dysarthria, stuttering, non-speech dysfluency, DNN, MFCC, LPCC

1. Introduction

Many studies have researched ways to improve the intelligibility of dysarthric speech, including methods that targeted particular aspects of speech to modify. Kain *et al.* [1] implemented a system of transformations that focused strictly on mapping vowels from individuals with dysarthria to vowels more characteristic of non-dysarthric speech. Those experiments showed an intelligibility increase of 6%. In 2013, Rudzicz [2] proposed a method that added the correction of other pronunciation errors and adjusted tempo. Among a cohort of listeners unfamiliar with the speech of people with cerebral palsy, word recognition rates increased by 19.6%. Crucially, the Levenshtein-based detection of phoneme repetitions and non-speech dysfluencies in that work depended on full phoneme segmentation, which may itself be quite challenging for dysarthric speech.

Chee *et al.* [3] provided an overview of automatic stuttering detection, emphasizing its difficulty across a number of classification methods. Czyzewski *et al.* [4], e.g., implemented artificial neural networks (ANNs) and 'rough sets' to detect three types of 'stuttering': stop-gaps, vowel prolongations, and syllable repetitions, obtaining accuracies up to 73.25% with ANNs and 91% with rough sets. Wiśniewski *et al.* [5, 6] performed two studies that used hidden Markov models with Mel-frequency cepstral coefficients (MFCCs) to detect stuttering. The first focused on both prolongation of fricative phonemes and blockades with repetition of stop phonemes that produced an accuracy of 70% [5]; the second strictly focused on prolongation of fricative phonemes and found an improvement in accuracy to approximately 80% [6].

Rath investigated modifications to MFCC feature vectors in speaker adaptation using deep neural networks (DNNs) [7], obtaining 3% improvements over Gaussian mixture models (GMMs) baselines. Across various types of speech features, deep learning has shown considerable improvements across several areas of speech recognition [8], compared with traditional techniques such as hidden Markov models. Here, we compare MFCCs (which are the most commonly used feature set in this domain [3]) and linear predictive cepstral coefficients (LPCCs), which are another popular but less utilized feature set. An exception was Chee *et al.* [9], who applied LPCCs with k-nearest-neighbors and linear discriminant analysis classifiers to automatically detect prolongations and repetition stutters, with recognition accuracy up to 89.77%. In the related field of automatic speech recognition ()ASR), MFCCs have consistently generated better results than LPCCs [10, 11]; to see if this trend extends to the domain of dysfluency detection, we compare these feature types with DNNs.

2. Methodology

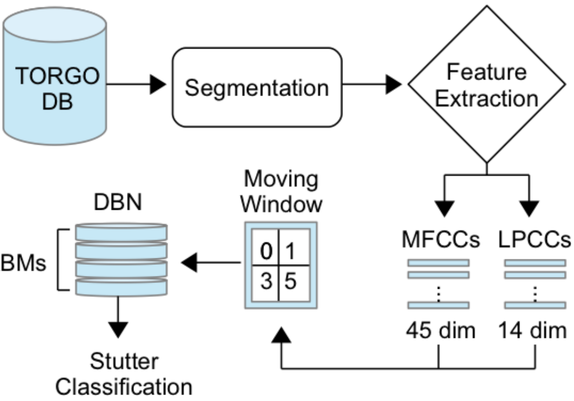


Figure 1: *Overview of automatic stuttering detection method.*

2.1. Data

The TORGO database [12] was created by a collaboration between the departments of Computer Science and Speech-

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 60–64,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

Language Pathology at the University of Toronto, and the Holland-Bloorview Kids Rehab hospital. The corpus consists of recordings from seven participants, three females and four males ranging in age from 16 to 50, diagnosed with cerebral palsy or amyotrophic lateral sclerosis. Additionally, there are recordings from seven control speakers matched for age and gender. A combination of non-words, short words, restricted sentences, and unrestricted sentences were recorded by all participants with a 16 kHz sampling frequency using two microphones. The database also includes articulatory measurements using electromagnetic articulography, which is not used here.

2.2. Segmentation

Segmentation was performed manually by listening to the recorded speech samples in the TORGO database and marking the start and end times of each occurrence of stutters. Only a single type of 'stuttering' dysfluency is considered here, specifically repetition-type stutters (Table 1), since these are more difficult to detect than prolongations and stop-gaps [4].

Table 1: *Repetition Types*

Repetition Type	Example
Part of a word	wh-wh-what time is it?
Whole word	what-what-what time is it?
Phrase	what time what time is it?

For the analysis of non-speech dysfluencies we employed the phonetic transcriptions provided with the TORGO database. In such transcriptions, non-phonetic segments are marked with the label *noi* (noise).

2.3. Feature extraction

After segmentation, speech data were parameterized into an input form suitable for use by a DNN classifier (Figure 2), as described below.

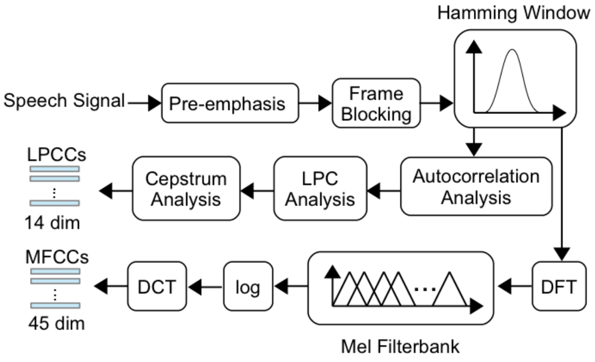


Figure 2: *MFCC and LPCC feature extraction overview.*

2.3.1. MFCC features

The MFCC input feature baseline consists of 13 cepstral coefficients in addition to the 0^{th} cepstral coefficient, energy, δ, and $\delta\delta$ coefficients. There is no pre-emphasis performed on these features. Since speech samples are constantly changing, we use frame blocking to analyze the signal in small time frames such that it becomes near stationary. The speech signals are cut into 25 ms frames with a frame step of 10 ms. We use a Hamming window to calculate the MFCC features, where the coefficients are found given Equation 1 (N is equal to window size minus one, in this case $N = 399$).

$$w(n) = 0.54 - 46cos(2\pi\frac{n}{N}), \qquad 0 \leq n \leq N \qquad (1)$$

To detect the different frequencies in the signal, the power spectrum is calculated using the discrete Fourier transform (DFT). The Mel filterbank then sums the energy in each filter, obtaining 29 uniformly-distributed triangular filters. The discrete cosine transform (DCT) is then applied to the log-filterbank energies to obtain the MFCCs. The purpose of the DCT is to decorrelate the overlapping filterbanks.

2.4. LPCC features

The LPCC features include 13 coefficients followed by the energy coefficient. LPCCs are more vulnerable to noise than MFCCs, so the speech signal is flattened before processing to avoid additive noise error. This is accomplished by pre-emphasis, a first order high-pass filter is applied to the speech signal as in

$$H(z) = 1 - az^{-1}, \qquad a = e^{-\frac{100\pi}{16000}} = 0.9806. \qquad (2)$$

Frame blocking and the Hamming window are applied to the LPCC feature space with the same parameters as for MFCCs (i.e., frame blocking 25 ms, 50% frame overlap and frame step 10 ms). This is followed by LPC analysis that estimates the coefficients by using the autocorrelation method to obtain fundamental frequency, pitch, and repeating patterns in the speech signal, before cepstral analysis is performed.

2.5. Feature modulation

We explored increasing the dimension of the input features used by the DNN due to the fact that DNNs are robust to larger input dimensions. The frequently-used hidden Markov model with Gaussian mixture output densities can become subject to error in parameter estimation, even with a slight increase in the input dimensions. The concept of a moving window is implemented to create inputs with larger dimensions. The moving window considers frames before and after the current frame. For example, a window of size $\pm x$ takes the x consecutive frames preceding and following the current frame and combines them into a single input vector (Figure 3 provides a visual representation of a moving window of size ± 1).

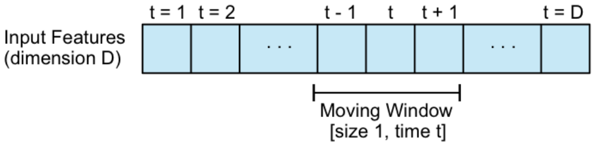


Figure 3: *Moving window of size ± 1.*

The dimensions of the input features are provided in Table 2. The baseline number of MFCCs and LPCCs are 45 and 14, respectively. The purpose of the moving window is to exploit the DNN's ability to use higher-dimensional input feature vectors to achieve better classification results by integrating contextual information.

Table 2: *Input feature dimensions*

Moving window size	0	±1	±3	±5
MFCC input dimension	45	135	315	495
LPCC input dimension	14	42	98	154

2.6. Classification

We use the deep neural network implementation of Tanaka and Okutomi [13] for stuttering classification. Four pre-trained Bernoulli-Bernoulli restricted Boltzman machines (RBMs) plus a decision layer are stacked to form a deep belief network (DBN), to create a DBN-DNN classifier (Figure 4). The RBMs are pre-trained in an unsupervised way using contrastive divergence. Once the DBN is initialized with the pre-trained RBMs, we fine-tune the DBN with a supervised learning method based on reducing error in the classification of, alternatively, stuttering or various types of non-speech dysfluencies.

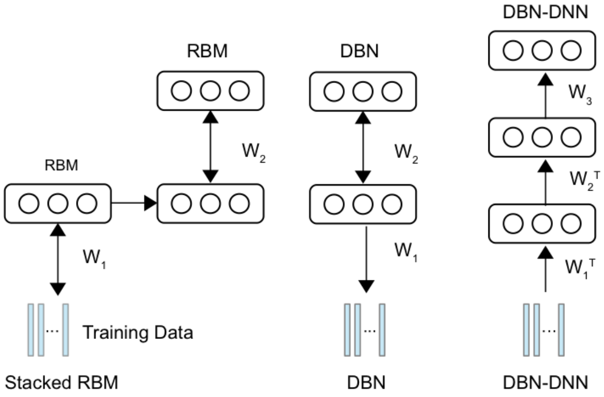


Figure 4: *DBN-DNN overview, after [13]*

3. Experiment 1: stuttering detection

We use two different partitioning schemes to compare results according to different categories of interest (Figure 5), namely generic-vs-individual speaker models (i.e., speaker-independent vs. speaker-dependent), and dysarthric-vs-non-dysarthric individuals. A total of 120 repetition stutters occurred across all 3115 recordings of dysarthric speech, and a total of 42 repetition stutters occurred across all 5641 recordings of non-dysarthric speech. The male and female dysarthric speakers with the most stutter occurrences were used for individual analysis; specifically, male dysarthric speaker M04 with 32 stutters, and female dysarthric speaker F03 with 22 stutters. Among the non-dysarthric speakers, there is no significant difference between males and females, so the non-dysarthric speaker with the most stutter occurrences was used in further analysis, namely male control MC04 with 16 stutters.

All training and testing data sets were divided in the same way – 70% of stutter occurrences were randomly assigned to training and paired with a random utterance without any stutter. By balancing training class sizes, we avoid the problem of overfitting to devolved majority classification. Testing data consisted of the remaining 30% of repetition stutters.

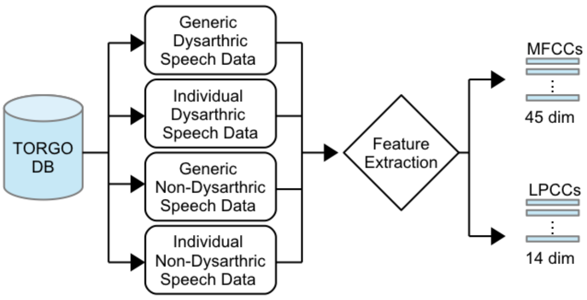

TORGO DB
Generic Dysarthric Speech Data
Individual Dysarthric Speech Data
Generic Non-Dysarthric Speech Data
Individual Non-Dysarthric Speech Data
Feature Extraction
MFCCs
45 dim
LPCCs
14 dim

Figure 5: *Training & testing data set divisions used in experimentation.*

An empirical question is whether stutter detection is more or less difficult in dysarthric speech, compared to non-dysarthric speech. Table 3 shows the average error rates of detecting repetition stuttering using 5-fold cross validation with MFCC and LPCC features. Clearly, across all models, accuracy increases monotonically as additional context is added. We also note that we obtain state-of-the-art accuracy for dysarthric speaker F03 using 10 frames of surrounding context, which is comparable to Czyzewski *et al.*'s work with rough sets [4]. An n-way analysis of variance reveals strong effects of window size ($F_3 = 836.91, p < 0.001$) and population ($F_1 = 11.80, p < 0.01$), but not of the feature set ($F_1 = 0.12, p = 0.74$). Across all experiments, LPCCs give slightly lower error than MFCCs, on average (20.17% vs. 20.32%, respectively). Except for the (relatively inaccurate) case where no context frames are used, generic control models always give higher error than generic dysarthric models, by absolute differences of 2% to 2.35%. It is important to note that we only consider main effects of these grouping variables – given the different dimensionality of MFCC and LPCC, one cannot make direct *interaction* comparisons across these groups and context sizes simultaneously.

Speaker-dependent models always outperformed associated speaker-independent models. The difference in error rates between generic and individualized models is larger for dysarthric speech than non-dysarthric speech. At best, the speaker-dependent dysarthric models achieved a 5.06% lower rate than the speaker-independent dysarthric models, while speaker-dependent non-dysarthric models obtained at best a difference of 2.85%.

Interestingly, it is easier to detect stuttering in dysarthric speech than in non-dysarthric speech. In fact, error rates were consistently lower for the dysarthric speech (≈14%) than for the non-dysarthric speech (≈16%). This suggests that the implemented method is robust to this particular speech disorder.

4. Experiment 2: non-speech dysfluencies

We repeated the methodology of Experiment 1, but considered instead 'lower-level' dysfluencies and non-speech vocal noise that can affect speech recognition and synthesis systems.

Here, annotation is based on the phonetic transcriptions provided in the TORGO corpus. Segments labeled as *noi* (noise) were examined and manually tagged with either none, or any combination of the following three dysfluency types:

aspiration Noise related to breathing, i.e., inspiration or expiration.

mouth/lips Noise produced by the lips and/or mouth/tongue.

vocal Non-speech voicing (e.g., laughter, hesitation...).

Table 3: *Average error rate (%, 5-fold cross-validation) of stutter detection using MFCC and LPCC features across speaker groups. Speakers F03, M04, and MC04 are also examined individually due to their relatively high rates of stuttering.*

		Window size			
	Speaker(s)	0	±1	±3	±5
MFCC	F03	38.36	9.93	9.74	9.55
	M04	38.61	12.80	12.70	12.60
	all dysarthric	40.84	14.95	14.82	14.61
	MC04	38.24	14.49	14.27	14.05
	all controls	40.00	17.30	17.09	16.88
	all speakers	40.74	15.21	15.07	14.93
LPCC	F03	38.31	9.87	9.50	9.13
	M04	38.56	12.81	12.61	12.41
	all dysarthric	40.80	14.95	14.68	14.42
	MC04	38.18	14.44	14.00	13.57
	all controls	39.94	17.26	16.84	16.42
	all speakers	40.70	15.20	14.92	14.64

The procedure of classification and evaluation is the same as in Experiment 1, except only individuals with dysarthria are considered, since the amount of occurrences of such dysfluencies in control speakers were not significant. Among all 1403 recordings of the head-worn microphones for dysarthric speakers with phonetic transcriptions, we found 706 instances of aspiration noise, 496 of mouth/lips, and 111 of vocal noise.

Table 4: *Average error rate (%, 5-fold cross-validation) across other dysfluencies using MFCC and LPCC features across speaker groups.*

		Window size			
	Type	0	±1	±3	±5
MFCC	aspiration	39.98	19.19	19.60	19.11
	mouth/lips	43.28	24.95	24.81	24.68
	vocal	46.15	25.75	26.83	25.81
LPCC	aspiration	40.08	19.35	19.40	19.14
	mouth/lips	43.31	25.01	25.03	24.83
	vocal	46.18	25.81	25.92	25.42

Table 4 shows the average error rates of detecting the different non-speech dysfluencies using 5-fold cross validation with MFCC and LPCC features. The accuracy increases with the use of one or more frames of context, but adding more than one frame does not improve the results. These types of low-level dysfluencies are significantly localized in time or highly characterized by their spectral shape. Therefore, adding more contextual information does not appear to improve classification.

Dysfluencies of type *aspiration* are consistently more accurately classified than *mouth/lips*, which in turn are easier to classify than *vocal*. The *aspiration* dysfluencies contain a very characteristic timbre which is easier to discriminate from other speech sounds than the other classes. On the other hand, *vocal* dysfluencies are the closest to actual speech phones, leading to a more difficult differentiation. We note that *aspiration* dysfluencies are usually longer and since, in our current setting, an entire region is tagged with the noise type without performing segmentation, frames containing *aspiration* may be systematically more accurately labelled than those with other more localized noises such as *mouth/lips* or *vocal*.

5. Discussion and future work

We investigated the ability of a DBN-DNN to classify repetition stuttering and non-speech dysfluencies in dysarthric and non-dysarthric speech using MFCCs and LPCCs as input. Results indicate that repetition stuttering is detected with very similar (though significantly different) error rates across dysarthric and non-dysarthric speech. Increasing the dimension of the input, across either feature to the DBN-DNN consistently lowers the error rate, and there is no statistically significant difference between using MFCC or LPCC input features. Moreover, we find that among non-speech dysfluencies, aspiration is more accurately identified than mouth/lip dysfluency, which in turn is more accurately identified than other vocal activity. In both cases, a greater investigation into the effect of context is needed.

Overall, the results achieved here are comparable to similar work discussed in Section 1. However, given the somewhat limited number of stuttering and non-speech disturbances within TORGO, the results can be considered preliminary; more work with additional data sets would be needed to make more conclusive claims.

Since dysarthric speakers are more likely to stutter than non-dysarthric speakers, this must be considered when comparing across groups, especially when comparing aggregate speaker-independent models. Future work includes additional types of stuttering detection, including prolongations and stop-gaps in spontaneous speech. We are also interested in extending and combining additional feature types, including autoencoders, and alternatives to the DBN structure itself. However, this paper has clearly shown that state-of-the-art stuttering detection, which had previously focused on non-pathological speech, can be applied to dysarthric speech. This automates a crucial component in systems that automatically improve the intelligibility of speech signals. Specifically, correcting dysfluencies has previously been shown to be a highly (if not the most) effective transformation that can be applied to speech signals [2]. Whereas that work depended on gold-standard phonemic transcriptions, our current work on stutters is relatively accurate given only the acoustics.

6. Acknowledgements

This work is partially funded by Thotra Incorporated, of which Frank Rudzicz is the CEO. It is also supported by an NSERC Discovery grant (RGPIN 435874) and a grant from the Nuance Foundation.

7. References

[1] A. B. Kain, J.-P. Hosom, X. Niu, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 2, pp. 743–759, 2007.

[2] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech and Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[3] L. S. Chee, O. C. Ai, and S. Yaacob, "Overview of automatic stuttering recognition system," in *Proc. International Conference on Man-Machine Systems*, no. October, Batu Ferringhi, Penang Malaysia, 2009, pp. 1–6.

[4] A. Czyzewski, A. Kaczmarek, and B. Kostek, "Intelligent Processing of Stuttered Speech," *Journal of Intelligent Information Systems*, vol. 21, no. 2, pp. 143–171, 2003.

[5] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smoka, and W. Suszyski, "Automatic detection of disorders in a continuous

speech with the hidden Markov models approach," in *Advances in Soft Computing*, 2007, vol. 45, pp. 445–453.

[6] ——, "Automatic detection of prolonged fricative phonemes with the hidden Markov models approach," *Journal of Medical Informatics & Technologies*, vol. 11, pp. 293–297, 2007.

[7] S. P. Rath, D. Povey, K. Vesel, and J. Cernock, "Improved feature processing for Deep Neural Networks," in *Intespeech*, 2013, pp. 109–113. [Online]. Available: http://www.danielpovey.com/files/2013_interspeech_nnet_lda.pdf

[8] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[9] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, "Automatic detection of prolongations and repetitions using LPCC," *International Conference for Technical Postgraduates 2009, TECHPOS 2009*, 2009.

[10] U. Bhattacharjee, "A comparative study of LPCC and MFCC features for the recognition of Assamese phonemes," *International Journal of Engineering Research & Technology*, vol. 2, no. 3, pp. 1–6, 2013.

[11] T. Gulzar, A. Singh, and S. Sharma, "Comparative Analysis of LPCC , MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks," *International Journal of Computer Applications*, vol. 101, no. 12, pp. 22–27, 2014.

[12] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[13] M. Tanaka and M. Okutomi, "A Novel Inference of a Restricted Boltzmann Machine," in *International COnference on Pattern Recognition*, 2014, pp. 1526–1531. [Online]. Available: http://www.ok.ctrl.titech.ac.jp/ mtanaka/ICPR2014mtanaka.pdf

Model adaptation and adaptive training for the recognition of dysarthric speech

Siddharth Sehgal[1], *Stuart Cunningham*[1,2]

[1]Department of Human Communication Sciences, University of Sheffield, Sheffield, United Kingdom
[2]Centre for Asssitive Technology and Connected Healthcare, University of Sheffield,
Sheffield, United Kingdom

s.sehgal@sheffield.ac.uk, s.cunningham@sheffield.ac.uk

Abstract

Dysarthria is a neurological speech disorder, which exhibits multi-fold disturbances in the speech production system of an individual and can have a detrimental effect on the speech output. In addition to the data sparseness problems, dysarthric speech is characterised by inconsistencies in the acoustic space making it extremely challenging to model. This paper investigates a variety of baseline speaker independent (SI) systems and its suitability for adaptation. The study also explores the usefulness of speaker adaptive training (SAT) for implicitly annihilating inter-speaker variations in a dysarthric corpus. The paper implements a hybrid MLLR-MAP based approach to adapt the SI and SAT systems. ALL the results reported uses UA-SPEECH dysarthric data. Our best adapted systems gave a significant absolute gain of 11.05% (20.42% relative) over the last published best result in the literature. A statistical analysis performed across various systems and its specific implementation in modelling different dysarthric severity sub-groups, showed that, SAT-adapted systems were more applicable to handle disfluencies of more severe speech and SI systems prepared from typical speech were more apt for modelling speech with low level of severity.

Index Terms: speech recognition, dysarthric speech, speaker adaptation, speaker adaptive training

1. Introduction

Dysarthria is the collective name for a group of motor speech disorders, which result from single or multiple lesions in the brain. It usually results in the loss of motor speech control due to muscular atrophy and incoordination [1, 2]. Across various aetiologies, dysarthric speech is usually characterised by *imprecise consonant production, reduced stress, slow speech rate, hypernasality, harsh and strained voice, muscular rigidity, spasticity, monopitch and limited range of speech movements* [1, 2]. Dysarthria can either be congenital, occurring with conditions such as in cerebral palsy, or acquired, where it develops due conditions such as a stroke or Parkinson's disease.

The effect on speech production of dysarthria is not limited to the musculoskeletal structures, but it can also affect parts of subglottal, laryngeal and supraglottal systems [3]. It usually leads to reduced intelligibility of speech, which can be inversely related to the severity of the underlying condition. On a broad operational scale, severity can be indexed as mild, moderate, severe or any approximation within, such as mild-moderate. For people with severe dysarthria, their speech can be largely unintelligible to unfamiliar listeners.

It is estimated that around 1% of UK population is diagnosed with a neurological disorder each year, although, not all the conditions lead to dysarthria. In UK alone; stroke (416 per 100,000), cerebral palsy (200-300 per 100,000) and Parkinson's disease (200 per 100,000) are amongst the most prevelant causes of motor speech disorders [4, 5].

1.1. Speech interface and dysarthria

Speech has provided an attractive interface for people with dysarthria by enhancing human-human & human-computer interaction. It can enable people with dysarthria to participate in social settings where they can interact with non-familiar communication partners. Moreover, speech as an interface can provide users with a more real-time communication experience to convey messages, in comparison to traditional hardwired switch based interfaces. Earlier studies have shown that systems that deploy automatic speech recognition (ASR) as an interface in a dysarthric setup can have a lower accuracy than hardwired switch-based systems, but, the final message transfer is around 2.5 times faster than the later, even with mis-recognitions followed by corrections [6, 7].

According to a report by [8], more than 70% of dysarthric population with Parkinson's disease or motor neuron disease and around 20% with cerebral palsy or stroke could benefit from some implementation of an augmentative or alternative communication (AAC) device. The benefits of such a setup has proved effective for dysarthric people using speech as an interface for natural communication [9] or enabling them to control physical devices through speech commands [7].

1.2. Automatic speech recognition for dysarthric speech

Dysarthric speech recognition has been investigated for more than two decades [10, 11]. The efficacy of commercial systems has been limited for speakers with mild or mild-moderate dysarthria [12, 13]. In general, decreasing recognition accuracy is linearly related to increasing severity. As a consequence, it has been concluded that the systems are not suited to the higher variability inherent in dysarthric speech.

From a research perspective; acoustic modelling, speaker adaptation and signal enhancement techniques have been explored by researchers to deal with variabilities and disfluencies in dysarthric speech.

The system can be (i) speaker dependent (SD) , which is modelled to recognise only a particular speaker, (ii) speaker independent (SI), which is a generic model map to recognise a range of seen and unseen speakers and (iii) speaker adapted (SA), which attempts to minimise the mismatch between a

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 65–71,
Dresden, Germany, 11 September, 2015.

generic baseline SI model and the intended target speaker. Both generative and discriminative techniques have been exploited to model the acoustics of dysarthric speech. Discriminative approaches like support vector machines has shown some level of success in small vocabulary tasks [14, 15], but by large continuous density HMMs (CDHMM) and its variants remain the most exploited and successful techniques used till date. To get robust model estimates for SD/SI systems, large amounts of training data is usually required. This is not practically viable, since dysarthric speech is afflicted with sparse and inconsistent data problems due to physical constraints, fatigue and muscular atrophy related to a specific individual. Moreover, any dysarthric system will only be effective in real time if the data is collected under conditions where the user will be engaged more often. To overcome this problem to some extent, researchers are using SA systems, which might give SD like performance using lesser amount of data and will be more apt for modelling any unseen user, if a good baseline SI model is available.

Earlier studies using CDHMMs suggested that speaker adapted (SA) systems were suited for mild to moderate dysarthric speakers and speaker dependent (SD) systems better modelled variablities in the severe group of speakers [13, 16]. However, till date there is no common consensus on an established scheme, which indicates the suitability of a technique for a specific type, aetiology or severity of dysarthria. For example, a study by [17], reported a contrary conclusion and suggested that severity is not a good indicator for an optimal selection of modelling approach. Their SA based system outperformed the SD system for most of the speakers used in the study. The disagreement over an optimal approach could also be due to (i) less number of speakers examined in a study, sometimes one, and, (ii) a small vocabulary size, which can create a bias for a certain technique due to the small homogeneous dataset.

1.3. Purpose and aim for the paper

There is a growing need to investigate SA based speech systems, which can be trained with less data and be more accurate for a reasonably large vocabulary. Preparation of SA system usually require using a baseline speaker independent (SI) system and then adapting it using standard techniques. The adaptation methods are usually model based, such as MAP [18] or applies a family of linear transforms, such as MLLR [19]. For dysarthric speech, the basline SI systems are usually prepared from a corpus of typical speech, dysarthric speech or a combination of both.

Although, little work has been done to investigate for an optimal adaptation approach, but some novel attempts have paved the path for further research and investigation. One of the earlier studies comparing SA and SD systems, was reported by [17]. The study was conducted for 7 speakers from the UA-SPEECH database [20] and the results showed that SA system outperformed the SD system for most of the speakers. A more comprehensive study was conducted by [21] on the same dataset that included all the speakers in the UA-SPEECH corpus. They tested a SD system alongside a MAP based SA system. An array of SI baseline models were used for adaptation purposes. Firstly the study showed an average relative increase of 34.5% over the earlier reported results by [17]. Secondly, the results showed that SI system using all the dysarthric speech data forms the best baseline system for MAP adaptation. To the best of our knowledge, the results reported by [21] seems to be the best till date on a relatively large vocabulary size of 255 words for a particular dysarthria type covering a range of severities.

This paper builds up upon these earlier studies and (i) investigates the best SI baseline system for adaptation of dysarthric speech, (ii) explores hybrid adaptation approach using MLLR-MAP and (iii) investigate the efficacy of speaker adaptive training (SAT) [22] to implicitly annihilate the inter-speaker variabilities during the training process.

In the paper, section 2 will detail about the data preparation and methodology used for the experiments, section 3 will present and analyse the recognition results, section 4 will put some collective discussion for the results and section 5 will have the concluding remarks and considerations for the future work.

2. Experimental Setup

2.1. Data preparation

All the experiments presented in this paper used two standard corpora for typical speech, viz., WSJ0 SI-84 [23] that consists of read speech from 84 North American english speakers with texts drawn from a machine-readable corpus of Wall Street Journal news, and, WSJCAM0 [24] , which is a British english version of WSJ database that consists of data from 92 training speakers. For WSJCAM0, data was also included for speakers from the development and two evaluation test sets.

In addition, UA-SPEECH [20] corpus was used, which consists of data from 15 dysarthric speakers with cerebral palsy and 13 control speakers. There are 765 isolated words (455 distinct) per speaker collected in three separate blocks, where each block consists of 10 digits, 26 international radio alphabets, 19 computer commands, 100 common words and 100 distinct uncommon words, which were not repeated across blocks. In addition, the corpus also provides a rough estimate of perceptual speech intelligibility ratings for each dysarthric speaker by five naive listeners. The ratings given will be used in all the experiments for ordering the speakers in various severity groups. All the block one (B1) and block three (B3) data from UA-SPEECH was used for training & adaptation purposes and block two (B2) was solely used for all the reported test results in the paper. Because dysarthric speakers can take a longer duration to utter words, the UA-SPEECH training data had to be logically re-segmented to get rid of extra silences around word boundaries. Only 200 ms of silence was appended to either side of the word for training. However, test data block B2 was left untouched to maintain the natural speaking conditions. Data from all the microphones was used for each corpus for training and adaptation purpose and a summary is given in Table 1.

Corpus	Speakers	Training Files
WSJ SI-84	84	14377
WSJCAM0 †	136	18537
UA-CTL	13	41819
UA-DYS	15	44277

Table 1: *A summary of each training corpus in the system. UA-CTL and UA-DYS codes are used for UA-SPEECH control and dysarthric speakers. (†) Four evaluation speakers with no secondary microphone data were excluded from WSJCAM0.*

For acoustic modelling, data from all the corpora was processed as 12 dimensional MFCC features with c_0 and cepstral mean normalisation. First and second order time derivatives were also appended giving a 39 dimensional feature vector per frame. Speech was analysed in 25 ms window with a 10 ms target shift rate.

2.2. Acoustic Modelling

The continuous density HMM in all the experiments are word-internal tied-state triphone models with clustering performed using phonetic decision trees. It follows a strict left-to-right topology with 16 Gaussian components used per state. Silence states were modelled using 32 Gaussian components.

2.3. Methodology

One of the aim of the paper is to test the efficacy of a good baseline SI system that is more apt for adaptation purposes. This is an extension of the SI systems that was described in [21]. Table 2 summarises the SI systems that were constructed for adaptation purposes.

System Code	Training Dataset Used
SI-00	WSJ SI-84 + WSJCAM0
SI-01	UA-DYS excluding target test speaker
SI-02	UA-DYS
SI-03	UA-CTL
SAT	UA-DYS

Table 2: *Summary of baseline systems and the corpus used for its preparation.*

The SI systems intrinsically model the speaker characteristics and acoustic realisations in speech, which are considered constant throught the database. During typical speaker adaptation, the optimal model set $\tilde{\Phi}$, given a set of S speakers in the system is generally represented as:

$$\tilde{\Phi} = \arg\max_{\phi} \mathcal{L}(O;\phi) = \arg\max_{\phi} \prod_{s=1}^{S} \mathcal{L}(O^{(s)};\phi)$$

where $\mathcal{L}(O^{(s)};\phi)$ is the likelihood of the observation sequences from speaker s, given the current set of model estimates ϕ.

In addition to various SI systems, SAT modelling was also considered in the current study, which splits information into various homogeneous blocks, e.g. data pertaining to a particular speaker for incorporating speaker induced variations. SAT training uses two sets of parameters, a canonical model ϕ_c, usually hypothesised to represent phonetically relevant speech variabilities, and the set of transforms $\mathcal{T}^{(s)}$ to represent the speaker variabilties. This is given as:

$$(\tilde{\Phi}_c, \tilde{\mathcal{T}}) = \arg\max_{(\phi_c,\mathcal{T})} \prod_{s=1}^{S} \mathcal{L}(O^{(s)};\mathcal{T}^{(s)}(\phi_c))$$

In the above equation speaker induced variations are modelled by $\mathcal{T}$ and the canonical model is updated, given each transform. The entire SAT paradigm works iteratively in an interleaved fashion and can be depicted as shown in figure 1.

SAT based on MLLR transforms should be able generate robust canonical model estimates, however, it comes with computational and memory overheads [25], making it impractical for implementation. Such issues are usually avoided by applying constrained MLLR (CMLLR) [26, 27], which uses the same transform for both means and variances. The transforms are computed for each homogeneous block of data. SAT with CMLLR results in a kind of feature normalisation during model training and have the same computational load as any other standard HMM update. Unlike SI models which can be directly used for recognition, SAT canonical model sets are not suited for direct decoding. Both systems are usually adapted to some target test condition.

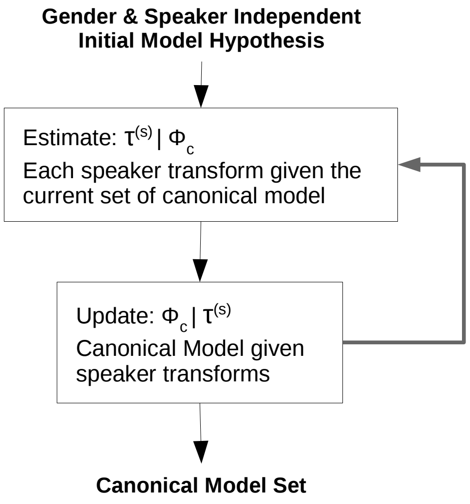

Gender & Speaker Independent
Initial Model Hypothesis
Estimate: $\tau^{(s)} | \Phi_c$
Each speaker transform given the current set of canonical model
Update: $\Phi_c | \tau^{(s)}$
Canonical Model given speaker transforms
Canonical Model Set

Figure 1: An overview of the SAT framework

In this paper, we present the results of the SI and SAT models using MLLR, MAP and MLLR-MAP based adaptation techniques. SAT canonical models are intentionally trained using only UA-DYS speakers to implicitly reduce the inter-speaker variabilities associated with dysarthric speech in general across varying degree of severities. The MLLR implemented uses a two-pass static adaptation procedure. The first pass performs a global transformation and the second pass uses the global transforms to produce more accurate transforms using a regression class tree with 32 terminal leaf nodes.

3. Results

All the test results presented in the paper are obtained on test set B2 of the UA-SPEECH corpus. Since the database comprises of single word utterances, the decoding grammar was strictly restricted to recognise only one of the possible test words, mostly preceded and succeeded by silences. There are 255 distinct competing words in the test block with a total of 22281 files from all speakers and microphones.

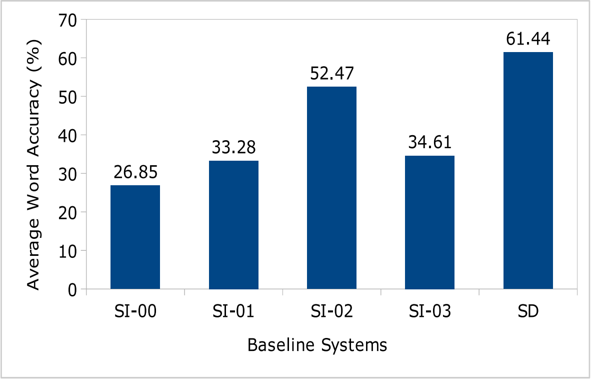


Figure 2: *Average word accuracy for the baseline SI systems along with the SD result.*

3.1. Baseline Systems

The first set of experiments involved obtaining recognition scores of all the baseline SI systems. These were then compared with the SD performance. Figure 2 shows the average baseline accuracy of all the SI systems. SI-00 has the lowest baseline result, which can be explained by the fact it was training only on typical speech. The high accuracy was obtained using the SI-02 system, which was trained on the largest amount of dysarthric speech data.

3.2. Baseline Adapted Systems

All of the baseline systems were adapted for each test speaker. Standard techniques were used and the results are shown in Figure 3. MAP clearly outperforms the MLLR based adaptation for all the systems except SI-00. This may be an example of non-informative priors. The SI-00 models are trained from WSJ0 + WSJCAM0 datasets, which contains only typical speech, and therefore presents no useful information about the model parameter distributions of the adaptation and test datasets.

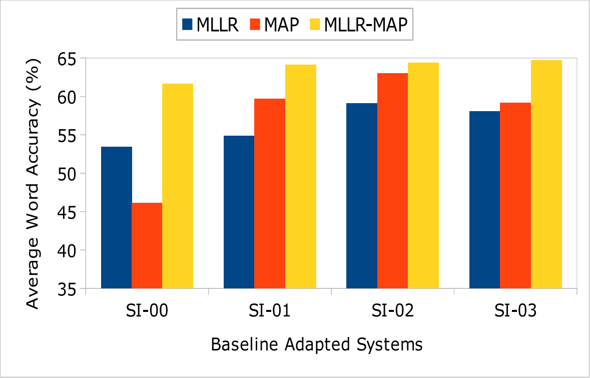


Figure 3: *Adaptation scores for the baseline SI systems.*

Following on from this observation we implemented a combined approach that involves generating MLLR transforms for the target speaker followed by MAP adaptation. By doing this, MLLR adapted parameters can act as informative priors for the MAP process. For all the SI systems, the MLLR-MAP combination outperformed all other adaptation approaches. For this reason the remainder of the paper will primarily focus on results obtained using a MLLR-MAP approach.

Intuitively, it may be thought that SI-01 or SI-02 should form an optimal set of baseline models for adaptation, since they exhibit less difference between the training, adapted and test conditions. Overall, the best MLLR-MAP scores for dysarthria and typical speech based SI systems was found to be for SI-02 and SI-03.

3.3. SAT-adapted vs Other Systems

One of the aims of the paper is to study the effect of SAT based modelling to reduce inter-speaker variations during training time. This section reports SAT-adapted results and compares it to the state-of-the-art SD system and other SI-adapted systems reported earlier. Figure 4 gives a comparison of the MLLR-MAP based SI and SAT systems. Clearly, SAT-adapted model sets outperform all the other tested systems

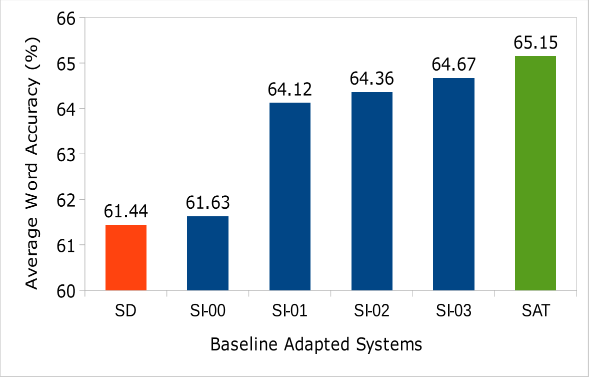


Figure 4: *Comparison of SD and MLLR-MAP based SI & SAT systems.*

It should be noted that SD system performs poorer than all the other adapted systems. Indeed, it can be seen in Table 3 that SD system does not perform better than any of the SA systems (*except one speaker*) under various intelligibility sub-groups. This gives us an average understanding that adaptation can be an effective approach to model dysarthric speech of varying severities. A similar finding about the efficacy of SA systems was also reported in a study by [17]. Our findings are contrary to some of the earlier published results [16, 13], which were more inclined to favour SD systems with increasing severity. In another study by [21], SI systems prepared from only dysarthric datasets produced better adapted models for most of the speakers.

In contrast our findings suggest that SI systems like SI-03, prepared from typical speech can also adapt as well as a dysarthric speech-based SI system. In order to justify our presumption, the effectiveness of all the MLLR-MAP based SAT and SI systems along with SD system was statistically analysed using Cochran's Q test. All the systems were tested for differences across all the test speakers. The null hypothesis was rejected at $\alpha = 0.01$, *degrees of freedom = 5*, which meant that all the systems were not equally effective for modelling dysartric speech in general. Later a pairwise Cochran's Q test was conducted between the system with the best absolute average score (SAT) and all others. The test showed that SAT was significantly different to all other systems at $p < 0.01$, except for the SI-03 system.

3.4. Severity Based System Results

So far we have reported all our findings averaged across all the test speakers regardless of the severity. However, to have a more customised approach for preparing systems for specific speakers it is important to individually study the effect of SD and SA based systems under various severity groups. The MLLR-MAP results reported earlier were investigated further for each of the different severity groups. Figure 5 gives an overall picture of how the baseline SI systems performed for various intelligibility sub-groups and Figure 6 shows the effect of adaptating the respective baseline systems along with SAT estimates. The speakers at the lowest intelligibility group showed inclination towards SAT based system or systems prepared with some dysarthric data, while, speakers in the highest intelligibility group benefitted from the presence of only typical speech data. Table 3 gives a detailed test report for all the UA-DYS speakers.

In order to understand differences between the systems, a Cochran's Q test was again applied to study the system differences under various speaker severity groups. The summary of the results of this test are shown in Table 4. It shows that SAT

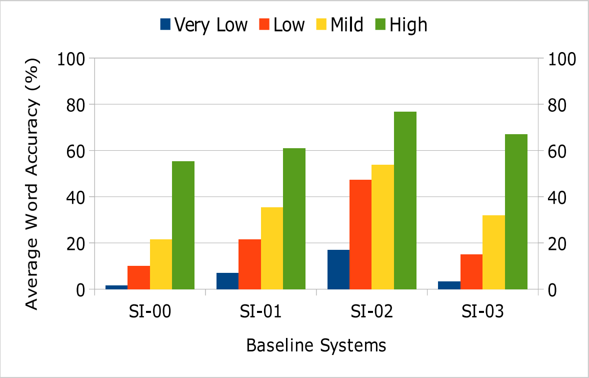


Figure 5: *Word accuracy for the baseline SI systems under various intelligibility groups (Very Low, Low, Mild, High).*

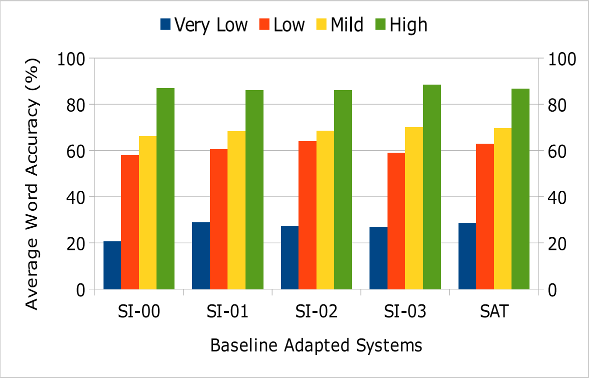


Figure 6: *MLLR-MAP scores for the SAT & SI systems under various intelligibility groups (Very Low, Low, Mild, High).*

Intelligibility	**Speaker**	**SD**	**MLLR-MAP**				
			SI-00	**SI-01**	**SI-02**	**SI-03**	**SAT**
Very Low	**M04 (2%)**	6.54	8.98	9.5	8.54	8.11	9.68
	F03 (6%)	32	27.61	37.49	36.01	36.81	38.36
	M12 (7%)	32.24	17.76	35.08	32.31	30.71	32.9
	M01 (17%)	16.76	27.03	28.32	28.22	27.46	29.22
Sub Acc.		23.52	20.61	28.82	27.36	26.95	28.71
Low	**M07 (28%)**	62.33	69.7	69.26	68.89	61.91	66.06
	F02 (29%)	61.08	37.62	50.12	54.02	50.93	56.93
	M16 (43%)	64.29	68.08	62.76	66.47	65.23	66.55
Sub Acc.		62.48	57.89	60.56	62.92	59.03	62.98
Mild	**M05 (58%)**	70.48	64.27	69.93	70.6	67.47	71.83
	M11 (62%)	58.18	56.57	63.8	66.06	68.1	65.62
	F04 (62%)	62.66	76.06	70.57	68.48	74.52	70.57
Sub Acc.		64.44	66.12	68.34	68.51	70.13	69.54
High	**M09 (86%)**	80.96	83.11	84.43	85.62	87.82	86
	M14 (90%)	77.76	80.4	80.09	79.2	85.71	80.84
	M10 (93%)	84.28	91.77	86.28	87.21	91.33	88.08
	M08 (95%)	85.86	87.96	87.21	86.47	87.4	87.34
	F05 (95%)	86.46	92.14	92.01	92.33	90.58	92.08
Sub Acc.		83.07	87.08	86.01	86.17	88.57	86.87
Overall Acc.		61.44	61.63	64.12	64.36	64.67	65.15

Table 3: *Average word accuracy rates for SD and all SI baseline systems adapted using MLLR-MAP. The table also shows sub accuracy scores under various intelligibility groups. The best scores are highlighted in grey for each row.*

system is statistically equivalent to some other systems in the *very-low, low and mild* sub-group of speakers.

Intelligibility	**Best performing systems** ($p < 0.05$)
Very Low	SAT, SI-01
Low	SAT, SD, SI-02
Mild	SAT, SI-03
High	SI-03

Table 4: *Cochran's Q analysis for all the systems under various intelligibility sub-groups.*

For the *high* intelligibility sub-group, system trained from typical speech data with similar recording and vocabulary setup as the test dysarthric conditions was significantly different to all the other competing systems.

4. Discussions

The results reported in Section 3 show that it is difficult to train a system to model the variabilities in dysarthric speech and to generalise to speakers of different severities. For example, when studying the performance of various baseline systems in section 3.1, it was interesting to note that SI-03 had similar performance to SI-01 system, despite being trained from typical speech data. We think that SI-03 models will be making use of information from homogeneous vocabulary and recording conditions as the test dysarthric conditions.

The findings also show that SD system were not the most effective to model dysarthric speech. This can be partially attributed to the relatively small amount of data per speaker in UA-SPEECH, especially when compared to previous studies in the literature [16, 13]. The test block B2 also comes with many unseen acoustic realisations in the form of 100 unique "uncommon words" and an SD system is usually only tuned to maximise the model fit for the seen data blocks during training. In contrast, a SA system might overcome this problem to some extent by using acoustic information present from other users in the baseline SI systems. This might be a contributing factor for all the adapted systems to be significantly better than SD system.

Another point of interest, reported in section 3.3, indicated that to model dysarthric speech in general, SAT and SI-03 systems were not significantly different. Hence the selection of a good baseline system to adapt from cannot depend on any particular dataset. It needs a more thorough investigation to understand the acoustics of dysarthric speech at an intra and inter speaker level. For instance, these results suggest that the variabilities in dysarthric speech can be better accommodated from modelling both typical and dysarthric domains. One such attempt was reported by [28], where background interpolation MAP was implemented to obtain an intermediate prior acoustic model to narrow the gap between two disparate SI systems (*typical & dysarthric*), albeit, the reported results were no better than those reported by [21]. Our best overall results, as reported in sections 3.3 & 3.4, are based on MLLR-MAP adapted SAT systems. It gives an absolute gain of 22.91% (54.36% relative) over results of [28] and an absolute gain of 11.05% (20.42% relative) over results of [21].

The choice of a particular system for a given target speaker is not completely clear, even when analysis is carried out at specific intelligibility levels. Table 4 indicates several possible choices in the lower intelligibility group of speakers. Since dysarthric speech will be more variable in the lower intelligibility group, the presence of SI-01 and SI-02 does not come in as a surprise as they will be inherently capable of modelling some of the common disfluencies. Although, the presence of SD system in the *low* intelligibility sub-group might suggest some corpus bias towards a particular speaker. It would appear that the choice of a baseline model for a particular target speaker may be determined by the amount of training data available.

Despite the fact that several alternatives appear to be equivalent for different groups of speakers, it is noticeable that SAT-based systems are among the best performing for the very low to mild groups of speakers. This may be due to the implicit capability of SAT to remove the speaker induced variations during training time. This speaker normalising might be having a nullifying effect on some complex variabilities present across all the speakers.

Among systems trained with typical speech, SI-03 is significantly a better base model for adaptation than SI-00. This is despite being trained with a smaller dataset. This may suggest that large quantities of typical speech data might not be necessary for the base models adapted to recognise dysarthric speech.

Lastly, as shown in Table 4, it is not surprising to observe that SI-03 was the best performing system for speakers with a high intelligibility. Perceptually, high intelligibility dysarthric speech is more akin to typical speech. Table 3 clearly shows the inclination of typical speech baseline systems (*SI-00, SI-03*) to model *high* intelligibility sub-group of speakers. In addition to acoustic similarities, as mentioned earlier, SI-03 system also has an additional benefit of homogeneous vocabulary and recording conditions.

5. Conclusions and future work

The current paper investigated the effectiveness of SAT-adapted, SD and SI-adapted systems to model dysarthric speech. We found that the hybrid MLLR-MAP based technique outperformed other adaptation procedures. All the MLLR-MAP based SAT and SI systems produced an absolute gain over similar results reported in earlier studies [21, 28] for this corpus. SAT-adapted systems had the highest overall average word accuracy for all dysarthric speakers. Although, systems trained from typical speech data with homogeneous recording conditions and vocabularies as the test dysarthric conditions were not significantly different to SAT-adapted systems.

It is difficult to assert at this time about the best strategy of SI or SAT based systems for robust adaptation and recognition of a target dysarthric speaker. SAT-adapted systems can implicitly model inter-speaker variabilities and proved to be significantly better at recognising speech from speakers with lower intelligibility. in contrast, typical speech systems were more inclined to model high intelligibility sub-group of speakers. The results also showed that that adaptation might be a better than corresponding SD systems to model dysarthric speech.

Despite the results reported here, there is still no consensus on the best approach to model dysarthric speech with varying severity, aetiology or type. Future work should investigate the SAT-based modelling approach, especially approaches for customising baseline systems prior to adaptation to a specific speaker.

6. Acknowledgements

This report is independent research funded by the National Institute for Health Research Invention for Innovation programme, Speech Therapy Apps for Rehabilitation, (II-LB-0214-20008). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

7. References

[1] F. Darley, A. Aronson, and J. Brown, "Clusters of deviant speech dimensions in the dysarthrias," *Journal of Speech and Hearing Research*, vol. 12, pp. 462–496, 1969.

[2] J. Duffy, *Motor Speech Disorders : Substrates, Differential Diagnosis, and Management*, 2nd ed. Elsevier Mosby, 2005.

[3] R. Kent, J. Kent, G. Weismer, and J. Duffy, "What dysarthria can tell us about the neural control of speech," *Journal of Phonetics*, vol. 28, no. 3, pp. 273–302, 2000.

[4] RCSLT, *Communicating Quality 3: RCSLT's Guidance on Best Practice in Service Organisation and Provision*. Royal College of Speech & Language Therapists, 2006. [Online]. Available: http://books.google.co.uk/books?id=udcuAAAACAAJ

[5] "Resource manual for commissioning and planning services for slcn," http://www.rcslt.org/speech_and_language_therapy/commissioning/aac_plus_intro, 2009, online; accessed on: 13-May-2015.

[6] M. S. Hawley, "Speech recognition as an input to electronic assistive technology," *The British Journal Of Occupational Therapy*, vol. 65, no. 1, pp. 15–20, 2002.

[7] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria." *Med Eng Phys*, vol. 29, no. 5, pp. 586–593, 2007.

[8] "Communication matters research matters: an aac evidence base," http://www.communicationmatters.org.uk/beyond-the-anecdote, 2013, online; accessed on: 13-May-2015.

[9] M. Hawley, S. Cunningham, P. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neill, "A voice-input voice-output communication aid for people with severe speech impairment," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 21, no. 1, pp. 23–31, 2013.

[10] M. Fried-Oken, "Voice recognition device as a computer interface for motor and speech impaired people," *Archives of Physical Medicine and Rehabilitation*, vol. 66, no. 10, pp. 678–681, 1985.

[11] C. Coleman and L. Meyers, "Computer recognition of the speech of adults with cerebral palsy and dysarthria," *Augmentative and Alternative Communication*, vol. 7, no. 1, pp. 34–42, 1991.

[12] K. Hux, J. Erickson, N. Manasse, and E. Lauritzen, "Accuracy of three speech recognition systems: Case study of dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, pp. 186–196, 2000.

[13] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *AAC: Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[14] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009.*, 2009, pp. 4605–4608.

[15] V. Wan and J. Carmichael, "Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, 2005, pp. 3321–3324.

[16] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, ser. Assets '07, 2007, pp. 255–256.

[17] H. Sharma and M. Hasegawa-Johnson, "State-transition interpolation and map adaptation for hmm-based dysarthric speech recognition," in *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, 2010, pp. 72–79.

[18] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[19] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[20] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 2008, pp. 1741–1744.

[21] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, vol. 2, 2012, pp. 1774–1777.

[22] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Fourth International Conference on Spoken Language, ICSLP 96., Proceedings.*, vol. 2, 1996, pp. 1137–1140.

[23] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, ser. HLT '91, 1992, pp. 357–362.

[24] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95.,*, vol. 1, 1995, pp. 81–84.

[25] M. Spyros, S. Rich, J. Hubert, and N. Long, "Practical implementations of speaker-adaptive training," in *DARPA Speech Recognition Workshop*, 1997.

[26] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[27] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[28] H. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech and Language*, vol. 27, no. 6, pp. 1147–1162, 2013.

Pronunciation Adaptation For Disordered Speech Recognition Using State-Specific Vectors of Phone-Cluster Adaptive Training

Sriranjani. R [*], S. Umesh[†] and M. Ramasubba Reddy [*]

[*]Biomedical Engineering Group, Department of Applied Mechanics
[†]Department of Electrical Engineering
Indian Institute of Technology - Madras
am12s036@smail.iitm.ac.in, umeshs@ee.iitm.ac.in, rsreddy@iitm.ac.in

Abstract

Pronunciation variation is a major problem in disordered speech recognition. This paper focus on handling the pronunciation variations in dysarthric speech by forming speaker-specific lexicons. A novel approach is proposed for identifying mispronunciations made by each dysarthric speaker, using state-specific vector (SSV) of phone-cluster adaptive training (Phone-CAT) acoustic model. SSV is low-dimensional vector estimated for each tied-state where each element in a vector denotes the weight of a particular monophone. The SSV indicates the pronounced phone using its dominant weight. This property of SSV is exploited in adapting the pronunciation of a particular dysarthric speaker using speaker-specific lexicons. Experimental validation on Nemours database showed an average relative improvement of 9% across all the speakers compared to the system built with canonical lexicon.

Index Terms: Dysarthric speech recognition, phone-CAT, lexical modeling, pronunciations, phone confusion matrix

1. Introduction

Clinical applications of speech technology play an important role in aiding communication for people with motor speech disorders. One such motor speech disorder is dysarthria, acquired secondary to stroke, traumatic brain injury, cerebral palsy etc. This affects more than one subsystem of speech production, leading to unintelligible speech. Some of the common characteristics of dysarthria include slurred speech, swallowing difficulty, slow speaking rate with increased effort to speak and muscle fatigue while speaking [1, 2]. All these effects affect the speech intelligibility but also the social interaction ability of people with speech disorders. Clinical applications of speech technology provide way to improve their communication in terms of the alternative and augmentative communication (AAC) devices. Automatic speech recognition (ASR) systems play a major role as an AAC device for aiding communication in terms of command/control in their daily lives. Only handful of databases are available for dysarthric speech, due to the fatigue and discomfort faced by the dysarthric speaker in providing data for longer time. With such constraints, acoustic models are usually built-in speaker adaptation framework [3, 4, 5].

The impairment in phonatory subsystem of a person affected with dysarthria leads to pronunciation errors. The slow rate of speech leads to a single syllable word being misrecognized as two syllable words. Frequent occurrences of non-speech sounds like hesitations false starts occur as part of dysarthric speech. These hesitations also lead to misrecognition of words as explained in [6, 4]. Imprecise consonant production is another characteristic of dysarthric speech. Since consonant production involves complex articulations compared to vowels, the errors are more frequent [7]. Muscle fatigue and lack of breath support increase the pronunciation errors of a dysarthric speaker [8].

All these effects increase the rate of insertions, substitutions, deletions and distortions in the dysarthric ASR systems. Thus the issue of pronunciation errors makes the design of dysarthric ASR system more challenging. The focus of this paper lies in handling these pronunciation errors especially substitutions by improving the lexical models. The lexicon contains the multiple pronunciations for each word expanded in terms of phones. The alternate pronunciations of a word is either formed manually [9] or obtained from the list of phone confusion pairs [10, 11]. This paper introduces a recently developed phone-cluster adaptive training (Phone-CAT) [12] acoustic modeling technique. Phone-CAT method build robust acoustic models using lesser number of parameters and limited amount of data. Thus the method can be used for limited data available domains especially in the case of dysarthric speech recognition. The main contributions of this paper are as follows:

- A novel approach to form speaker-specific phone confusion matrix using the low-dimensional SSV of Phone-CAT
- Using the speaker-specific phone confusion matrix to identify the confusion pairs (substitution phones) to form alternate pronunciations in the speaker-specific lexicon

Our proposed approach helps in forming phone confusion matrix directly from the Phone-CAT acoustic model, compared to the existing methods [10, 11] which align the decoded transcription with canonical transcription to form the phone confusion matrix. Thus we circumvent the usage of expensive decoding process. This preliminary study using Nemours database shows a relative performance improvement of 9% using our proposed approach compared to baseline model built using canonical lexicon.

2. Related work

Multiple pronunciations of a word in the lexicon improves the recognition performance. The lexical models are improved either implicitly or explicitly handling the pronunciation errors [13]. In order to improve the lexical models, the phones mispronounced by each dysarthric speaker need to be identified. Earlier work handled multiple pronunciations using expert knowledge by adapting pronunciations manually [9]. Per-

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 72–78,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

sonalized speaker articulation patterns were obtained from the speaker-adapted models along with the confusion matrix. These speaker-adapted models were obtained using universal disordered matrix and the posterior probability from the ASR system in an unsupervised fashion [13].

Another approach for identifying the mispronounced phones is by aligning the decoded text with the true transcription. A phone confusion matrix is formed using the decoded transcription and canonical transcription. This phone confusion matrix is used to identify the mispronunciations [10]. The substitution, insertion and deletion errors, were modeled as discrete hidden Markov model (HMM) called metamodels [11]. Another variant of this system is to train the extended metamodels from an integrated confusion matrix using genetic algorithm [14].

The concept of weighted finite state transducer (WFST) improves the performance of speech recognition systems. Composing confusion matrix along with the lexicon and language models in the WFST framework provides complementary information to the system. This concept was used in speech recognition [15] and keyword searching [16]. In dysarthric speech recognition framework, different methods were used to form confusion matrices to be used with WFST. One such method is to use KL distance measure between two context-dependent triphones to form confusion matrix [17, 18]. Deep neural networks (DNN) can also be used to improve pronunciation models. The posterior probabilities from pre-trained DNN were used to identify mispronunciations. They were further analyzed to generate pronunciations to form speaker-specific lexicons [19]. All the above methods, uses confusion matrix obtained by aligning the decoded transcriptions with the canonical transcriptions to improve the lexicons.

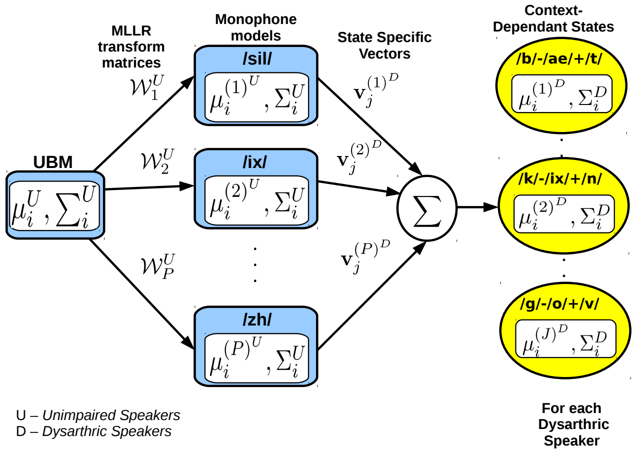


Figure 1: Phone-CAT architecture

In this paper, a novel approach is proposed to form the confusion matrix using the low-dimensional vector from the Phone-CAT acoustic model. Each tied-state in Phone-CAT model is modeled using SSV. The dominant weight of the SSV represents the pronounced phone. The mispronounced phone of each dysarthric speaker obtained using SSV is compared with the canonical phone to form the phone confusion matrix. This matrix is used to improve the lexical models by providing alternate pronunciations of words. Since each speaker has a separate pronunciation pattern, speaker-specific lexicons are formed. Acoustic models rebuilt using these lexicon, improve the performance of the system.

3. Phone-Cluster Adaptive Training Acoustic Models

The acoustic models are usually built using hidden Markov model–Gaussian mixture model (HMM–GMM) framework. The acoustic variations of speech due to age, gender, environmental changes and pronunciation variations are being modeled using GMM. The sequence information involving co-articulation is modeled as HMM. The triphone model represents a phone along with its left and right contexts capturing the co-articulation effects. For example, consider the triphone $/ax/ - /b/ + /k/$ representing the model for the center phone $/b/$, capturing the effect of its left context $/ax/$ and right context $/k/$. Several triphones with similar acoustic characteristics and same center phone $/b/$ are clustered to form a single tied-state. The GMM parameters are then estimated independently to model each tied-state. This estimation requires huge number of parameters and sufficient amount of data. This issue is handled using the recently proposed phone-CAT acoustic model by robustly modeling the available data with lesser number of parameters.

Phone-CAT is a HMM-GMM system in which the GMM parameters are represented in a compact form. In other words, the GMM for each tied-state is formed by the linear combination of all the monophone GMMs in that language. For example, the tied-state $/ax/ - /b/ + /k/$ containing triphones $/ax/-/b/+/k/$,$/ch/-/b/+/k/$,$/ae/-/b/+/k/$ is formed from the linear combination of all the monophone GMMs like $/sil/, /ax/, .../k/, ..., /zh/$. The weights of each monophone GMMs are represented by $v_j^{(1)}, v_j^{(2)} \dots v_j^{(P)}$, where P is the number of monophones. The vector containing the monophone weights is called SSV and is represented as $\mathbf{v}_j = \begin{bmatrix} v_j{}^{(1)} & v_j{}^{(2)} & . & . & v_j{}^{(P)} \end{bmatrix}^T$ with P dimensions. The monophone GMMs are in turn formed by adapting the universal background model (UBM) using maximum likelihood linear regression [20] transformation. The UBM is a GMM built using the available speech data from all the speakers. This UBM is adapted using the transformation matrices $\boldsymbol{\mathcal{W}}_1, \boldsymbol{\mathcal{W}}_2, ..., \boldsymbol{\mathcal{W}}_P$ for each of the P monophones, forming P monophone GMMs. The Phone-CAT architecture is shown in figure 1. The GMM parameters of the tied-state model are: means $\boldsymbol{\mu}_{ji}$, covariances $\boldsymbol{\Sigma}_i$ and Gaussian priors w_{ji}.

The mean parameter for each monophone models $\mu_i^{(p)}$ with Gaussian mixture i is combined to form the mean parameter of the tied-state j using the following equations:

$$\mu_i^{(p)} = \boldsymbol{\mathcal{W}}_P \boldsymbol{\xi}_i = \boldsymbol{\mathcal{W}}_P [\mu_i \quad 1]^T$$

$$\boldsymbol{\mu}_{ji} = \sum_{p=1}^{P} \mu_i^{(p)} \, v_j^{(p)}$$

Here $\boldsymbol{\xi}_i$ is the extended mean vector $[\mu_i \quad 1]^T$ with μ_i as the canonical mean of the Gaussian component i of the UBM. Since the mean $\boldsymbol{\mu}_{ji}$ and the Gaussian prior w_{ji} are represented in terms of the vector SSV $\mathbf{v}_j$ as in [12], the parameters are represented in low dimensions. Also the covariances $\boldsymbol{\Sigma}_i$ are estimated in a shared fashion across the tied-states. This reduction in the number of parameters helps in reducing the amount of data needed for estimation. More details of the model training and estimation of each parameters are explained in [12].

4. Importance of state-specific vectors

The SSV is a low-dimensional vector of dimension P representing each tied-state j uniquely. It captures the context information since it represent the weights with which each monophone GMM linearly combine to form a single tied-state. We know that different triphones with the same acoustical characteristics are tied together in order to form tied-state. The SSV plot of the second state of the triphone $/ch/ - /ix/ + /ng/$ is shown in figure 2.

It is clearly shown that, the dominant weight corresponds to the center phone $/ix/$. Apart from the center phone, the left and right context phones also get some considerable weight. The negative value represents the direction of the vector, but we are interested only in the absolute value of the elements of the SSV.

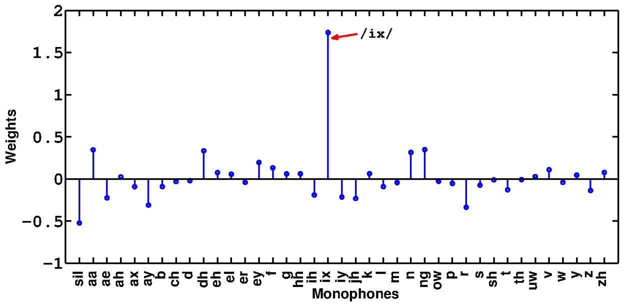


Figure 2: SSV plot of the second state of the triphone $/ch/ - /ix/ + /ng/$

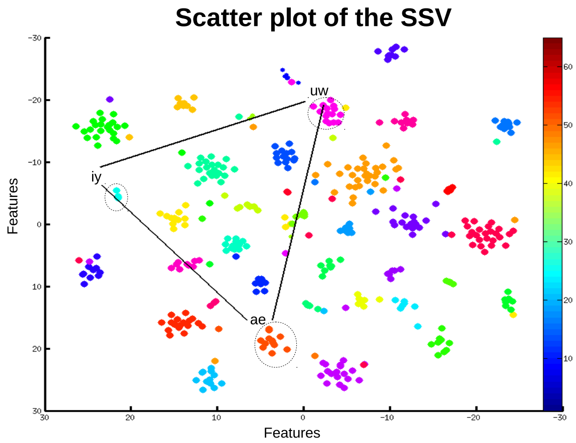


Figure 3: This two-dimensional scatter plot is obtained using the t-SNE toolkit by plotting the SSV of all the tied-states in Nemours database

A statistics of the dominant weight property of SSV was performed for unimpaired (control) speech data from Nemours speech database. The aim of the task was to check the statistics of the SSV picking the center phone of the tied-state correctly. It was found that out of 204 tied-states, the dominant component of SSV correctly picks the center phone 76% of the times and the top three weight values in the SSV picks up center phone 88% of the time. A similar analysis was also performed for the standard Switchboard database (≈300 hours of data), with 2400 tied-states. It was found that 70% of the time, the center phone was correctly picked up by the dominant component of SSV. Also ≈92% of the time, the left/center/right phones are picked up as the dominant component of SSV [21]. This shows that the SSV uniquely represents the enunciated phone (center phone of the tied-state) through its dominant weight most likely. The scatter plot of the P dimensional SSV reduced to two dimension is shown in the figure 3. The SSV related to each cluster represents a particular monophone (each in different color, a total of 39 phones were present in the Nemours database). These clusters are located at articulatory position of the vowel triangle in a well discriminated manner. This shows that SSV has the capacity to capture the phonetic information along with context information. Thus the analysis of SSV in this section leads us to the following conclusions:

- The dominant weight in SSV most likely represents the enunciated phone (center phone) of the tied-state
- Provides discriminable phonetic class information, since each vector is modeled for a particular tied-state
- SSV is hypothesized to capture the pronunciations of each dysarthric speaker when speaker-specific Phone-CAT models are built

This leads us to proceed to the proposed method of building Phone-CAT model specific to each speaker, thereby capturing the pronunciations of each dysarthric speaker.

Table 1: Extract dysarthric enunciated phone from SSV

Tied-states	**Phones**						
(Canl)	$\boldsymbol{sil}$	$\boldsymbol{aa}$	$\ldots$	$\boldsymbol{ey}$	$\ldots$	$\boldsymbol{zh}$	**Dysp**
$* - /\mathbf{sil}/ + *$	$\boxed{\mathbf{1.75}}$	0.21	$\ldots$	0.38	$\ldots$	0.12	$/sil/$
$* - /\mathbf{aa}/ + *$	0.03	$\boxed{\mathbf{0.09}}$	$\ldots$	0.01	$\ldots$	0.08	$/aa/$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$* - /\mathbf{jh}/ + *$	0.19	0.11	$\ldots$	$\boxed{\mathbf{0.36}}$ $\ldots$		0.13	$/ey/$
$ch - /\mathbf{ix}/ + ng$	0.48	$\boxed{\mathbf{0.50}}$	$\ldots$	0.01	$\ldots$	0.22	$/aa/$
$n - /\mathbf{ix}/ + k$	0.10	$\boxed{\mathbf{0.90}}$	$\ldots$	0.25	$\ldots$	0.76	$/aa/$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$* - /\mathbf{zh}/ + *$	2.01	0.02	$\ldots$	0.10	$\ldots$	$\boxed{\mathbf{3.06}}$	$/zh/$
Canl - canonical pronunciation; Dysp - dysarthric pronunciations							

The numbers inside circle shows the absolute maximum value in each SSV corresponding to dysarthric pronounced phone

5. Proposed Method for Improving Lexical Models

5.1. Phone-CAT model for each dysarthric speaker

The major step of our proposed method is to build speaker-specific Phone-CAT model. Initially, using the unimpaired speaker's data in the dysarthric database, a Phone-CAT model is built. The speaker-specific Phone-CAT model is obtained from the unimpaired speaker model by re-estimating the SSV and providing dysarthric speaker's data in maximum likelihood framework. The SSVs are initialized as (1/number of monophones), to allow the system to learn the weights of the monophone GMM using the available dysarthric speaker's data. At the end of this training process, Phone-CAT speaker-specific models are built. The architecture of speaker-specific Phone-CAT model is shown in figure 1. Finally, we obtain a set of tied-states specific to each dysarthric speaker from the speaker-specific Phone-CAT model.

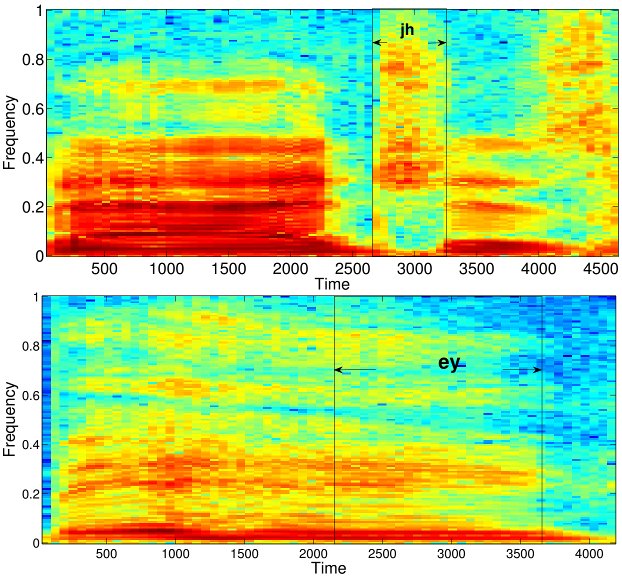


Figure 4: Spectrogram of the word "Badge" spoken by unimpaired speaker (on top) and dysarthric (BV) speaker (bottom). The spectrograms are plotted for a part of the waveform containing "The Badge is lifting the Beige".

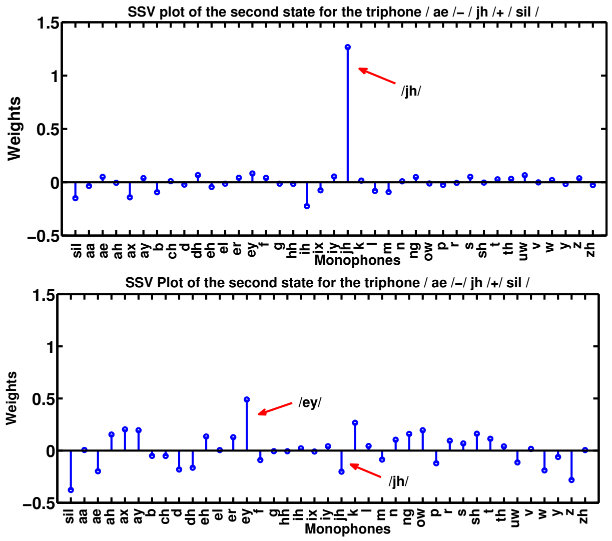


Figure 5: SSV for the second state of the triphone $/ae/-/jh/+/sil/$ for the unimpaired speaker (top) and dysarthric speaker BV (bottom).

5.2. Identification of mispronunciations made by the dysarthric speaker using SSV

Having built the speaker-specific Phone-CAT models, the next step is to extract the unique SSV associated with the set of tied-states. The P-dimensional SSV is extracted from each dysarthric speaker's Phone-CAT model. Using the dominant weight property of SSV discussed in section 4, the absolute maximum weight value of the SSV is obtained for each tied-state from each dysarthric speaker's Phone-CAT model. Since the pronounced phone is captured by the dominant weight of the SSV, the phone corresponding to the absolute maximum weight is hypothesized as the pronunciations made by the dysarthric speaker as shown in table 1. There may be cases where the canonical pronunciation (center phone of the tied-state) does not represent the observed pronunciation (phone associated with the absolute maximum weight of the SSV). In that case, it means that the phone model built for the speaker represents the observed pronounced phone rather than the canonical pronounced phone.

5.2.1. Analysis of the mispronunciations picked up by the SSV

Figure 5 shows the SSV plot for the second state of the triphone $/ae/ - /jh/ + /sil/$ of unimpaired and dysarthric speaker (BV). As discussed in Section 4, the dominant weight indicates the pronounced phone $/jh/$ for the triphone of the unimpaired speaker. But for dysarthric speaker, instead of $/jh/$, the phone $/ey/$ gets the dominant weight. This indicates that the phone $/jh/$ is mispronounced as $/ey/$. In order to analyze our hypothesis of the dysarthric speaker pronunciations captured by the dominant weight of the SSV, perceptual test was conducted. The audio samples of the BV speaker in the context for the word "$Badge - b\ ae\ jh$" (canonical pronunciation) was heard as "$Badge - b\,ae\,ey$" (dysarthric speaker's observed pronunciation). The audio sample was verified by 10 naive listeners and their mean opinion score was taken.

To further support this hypothesis, the spectrogram of the word "$Badge$" pronounced by unimpaired and dysarthric speaker is shown in figure 4. The fricative $/jh/$ is clearly visible in the spectrogram of unimpaired speaker, while for dysarthric speaker the diphthong $/ey/$ occurs instead of actual phone $/jh/$. Hence the second state of the triphone model $/ae/-/jh/+/sil/$ is more acoustically closer and represents the second state of the triphone model $/ae/-/ey/+/sil/$, captured directly by the SSV using its dominant weight. Thus we confirm our hypothesis that the phone captured by the SSV using its dominant weight corresponds to the pronunciations made by the dysarthric speaker.

5.3. Formation of phone confusion matrix using SSV

Using the SSV corresponding to a tied-state for a particular dysarthric speaker's model, a phone confusion matrix is formed. The set of canonical pronunciations (center phone of the tied-state) and the set of observed dysarthric speaker's pronunciations (absolute maximum weight of the SSV for each tied-state) are used to form the phone confusion matrix. From each dysarthric speaker's Phone-CAT model, speaker-specific phone confusion matrix is formed. Each row of the matrix corresponds to canonical pronunciations and each column represents the observed dysarthric speaker's pronunciations. The sum of all elements of the matrix corresponds to the total number of tied-states.

The diagonal elements represent the number of correct pronunciations made by the speaker, where the center phone of the tied-state is correctly picked up by the SSV as its dominant weight. The off-diagonal elements represents the mispronunciations made by that speaker, where the center phone does not correspond to the dominant weight of the SSV. Value in each element of the matrix say aij, represents the frequency of occurrence of the canonical phone i being mispronounced as phone j. This phone confusion matrix also correlates with the intelligibility scores of the different severity levels of the speakers. Since the diagonal elements represent the correct pronunciations, the number of elements across the diagonal varies with respect to the severity level of dysarthria. As the degree of impairment increases, the diagonal pattern disintegrates. Thus phone confusion matrix helps in the objective assessment of dysarthric speech [22]. Apart from assessment, phone confusion matrix is used for improving the lexical models which is the main focus of this paper.

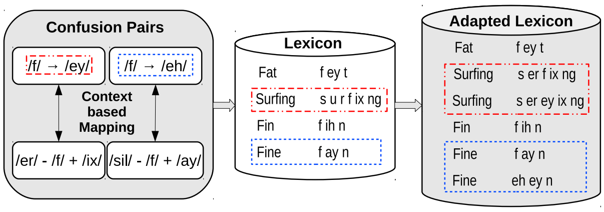


Figure 6: Schematic diagram of the proposed method to build an adapted lexicon from the phone confusion pairs using context dependent mapping. Here the canonical phone /f/ is confused with /ey/ when /f/ occurs in between the context /er/ and /ix/. Hence the word "surfing — s er **f** ix ng" gets the alternate pronunciation as "surfing — s er **ey** ix ng" while the words fat, fin and fine are neglected.

Algorithm 1 **Procedure to form phone confusion matrix and modified lexicon from SSV**

```
1. For each dysarthric speaker
    (a) Build the Phone-CAT acoustic model and extract the P-dimensional SSV from the set of tied-states
    (b) Absolute maximum weight of each SSV is picked up as the dysarthric speaker's pronounced phones as shown in table 1
    (c) Using the set of canonical pronunciations (center phones of the tied-states) and the observed dysarthric pronunciations (absolute maximum values of the SSV), a phone confusion matrix is formed
    (d) The list of substitution phones are obtained from the phone confusion matrix using a threshold
    (e) Obtained confusion pairs are mapped only for those canonical phones when their context matches with the corresponding tied-state
    (f) The modified lexicon along with the alternate pronunciations is then used to rebuild the acoustic models
```

5.4. Improved lexical models using SSV

The phone confusion matrix captures the mispronunciations made by each dysarthric speaker. The list of substitution phones are obtained from the set of mispronunciations in the matrix using a threshold rule. The substitution phones are further used to form alternate pronunciations forming speaker-specific lexicons. For example, if the phone $/f/$ is mispronounced as $/ey/$ in the confusion matrix with high recurrence, then it is taken as substitution phone. For the word "$five$" the alternate pronunciation in the lexicon is given as:

$$[Five]- > /\mathbf{f}\ ay\ v/\ \ \textit{(canonical pronunciation)}$$

$$[Five]- > /\mathbf{ey}\ ay\ v/\ \ \textit{(alternate pronunciation)}$$

It was shown that adding context-dependent pronunciation variation models helps in improving the performance of the system [18]. The triphones corresponding to the phone confusion pairs, are used to substitute the phones in the lexicon, for the words with the corresponding triphone context information as in figure 6. In the figure, $/f/$ is substituted with $/ey/$ only for the word "surfing" which contains the triphone context $/er/-/f/+/ix/$. For other words with phone $/f/$, no alternate pronunciations were given. This helps in reducing the size of the lexicon.

The number of confusion pair to be substituted from the confusion matrix is chosen based on the threshold rule. This helps in reducing the number of confusion pairs avoiding the selection of alternative confusion pair for each canonical phone. This modified lexicon is composed with grammar in WFST framework. In this approach, we mainly focus on modeling the substitution errors using the alternate pronunciations. Further, the modified lexicon is used to rebuild the acoustic model in the HMM-GMM framework.

6. Experimental setup

The experiments were performed in Kaldi [23] open-source speech recognition toolkit. Nemours database [24] was used for our experiments. It contains continuous speech utterances with 16 KHz sampling rate. It has 11 speakers, out of which only 10 speakers were used for our experiments [24]. Each speaker recorded 74 nonsensical sentences of the form "The N1 is Ving the N2" where the N1 and N2 are monosyllabic noun and V is the disyllabic verb. The lexicon is expanded in terms of phones with vocabulary size of 113 words and 39 phones in ARPAbet (advanced research project agency) symbol set is used for experimentation. One unimpaired speaker's data covering all the sentences spoken by each dysarthric speaker was recorded as control subject. The standard Frenchay dysarthric assessment (FDA) scores were also provided for each dysarthric speaker. The train data contains 490 utterances and test data contains 250 utterances, selected using 3-fold cross validation procedure. Trigram language model was used and the performance of the continuous density hidden Markov model (CDHMM) is measured using word error rate (WER). Baseline CDHMM is built with 200 tied-states and 10 Gaussian mixture components. The baseline system uses the canonical lexicon for both training and testing.

Rate	Mdl	FB	MH	BB	LL	JF	RL	RK	BK	BV	SC
Ins	Base	0	0	1	0	2	6	0	8	0	7
	Expt 1	0	0	1	0	2	6	0	7	0	7
	Expt 2	0	0	0	0	0	5	0	6	0	3
Del	Base	0	0	0	0	0	0	4	18	0	0
	Expt 1	0	0	0	0	0	0	2	17	0	0
	Expt 2	0	0	0	0	0	0	0	17	0	0

Table 2: Rate of insertions (Ins) and deletions (Del) for different models (mdl): Baseline (Base), Expt 1 and Expt 2

7. Results and Discussion

7.1. Results with modified lexicons : Proposed method

Two different experiments were performed to compare with the baseline CDHMM (Base) system. First is to train acoustic model using canonical lexicon and decoding the text using the modified lexicon (Expt 1). The second experiment is to use the modified lexicon for both training and testing process (Expt 2). Speaker-wise results for both the experiments are shown in the table 4. Comparing with baseline, all the speakers obtain improved performance for the system rebuilt using the modified lexicons (Expt 2). On an average, the relative improvement is 13.1%. Comparing with baseline system, an relative improvement of 5.4% is obtained across all the speakers for the system only tested using modified lexicon (Expt 1). Severe category speakers shows considerable improvement compared to moderate and mild category speakers. Substitutions form a major portion of the error compared to insertions and deletions in our model. Hence we focused on reducing the number of substitution errors in this paper. Figure 7 shows the reduction in the rate of substitutions for the proposed model compared to baseline system. The rate of insertions and deletions were also reduced which is shown in table 2.

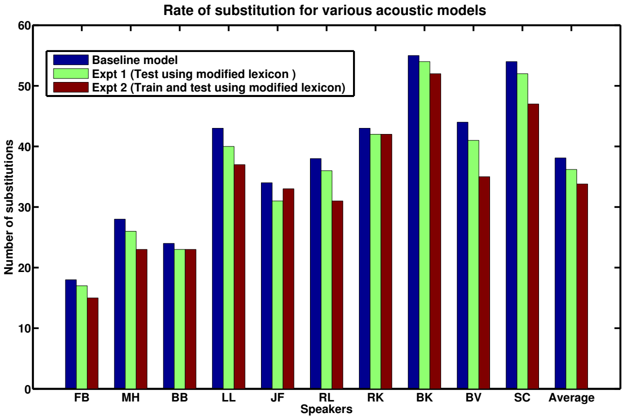


Figure 7: Substitution rate for different models

Table 3: Comparison of proposed method with existing method

Lexicon Usage	Method/ Model type	MH (Mild)	RK (Moderate)	SC (Severe)
Training $canonical$ + Testing $canonical$	Baseline	18.7	31.3	29.3
Training $canonical$ + Testing $modified$ (Expt 1)	Existing method	17.3	29.3	27.3
	Proposed method	16.7	29.3	27.3
Training $modified$ + Testing $modified$ (Expt 2)	Existing method	16.0	30.0	26.0
	Proposed method	15.3	29.3	23.3
	% R.I	4.2	2.2	10.3

Here % R.I denotes relative improvement with respect to existing method

7.2. Comparison with Existing Method

Some of the existing methods in literature involve forming phone confusion matrix aligning the recognized phoneme sequence with reference transcriptions [10]. Then using rule-based method, speaker dependent multiple pronunciation lexicons are formed. In [11], the recognized transcription is aligned with the reference transcription to form the phone confusion matrix. The confusion pairs are then used to provide multiple pronunciations in the lexicon. In order to compare our proposed method with the existing method, the confusion pairs from the phone confusion matrix formed by aligning the recognized transcription with the reference transcription on baseline model are used to form the lexicon.

Table 4: Results of lexical modeling for Nemours database in terms of % word error rate (% WER)

Severity	Speakers	Baseline CDHMM	Testing using new lexicon (Expt 1)	Train + Test using new lexicon (Expt 2)
Mild	FB	12.0	11.3	10.0
	MH	18.6	17.3	15.3
	BB	16.6	16.0	15.3
	LL	28.6	26.6	24.6
Moderate	JF	24.0	22.0	22.0
	RL	29.3	28.0	24.0
	RK	31.3	29.3	29.3
Severe	BK	54.0	52.0	50.0
	BV	29.3	27.3	23.3
	SC	40.6	39.3	33.3
Average		**28.7**	**26.9**	**24.7**

Similar to section 7.1, two different experiments (Expt 1 and Expt 2) were performed on baseline model using the modified lexicon formed using this phone confusion matrix for three different severity category. As shown in table 3, proposed method using phone confusion matrix formed from SSV shows an relative improvement of 10.3% compared to existing method using phone confusion matrix formed using decoded transcription. In the existing method, a single frame is involved in estimating the likelihood with respect to the corresponding acoustic model. While in case of our proposed approach, a set of frames corresponding to a tied-state label is involved in the estimation of SSV. Thus the estimated SSV are more reliable in identifying the confusion pairs which helps in improving the recognition performance over existing method.

8. Conclusions

This paper focuses on improving the performance of dysarthric speech recognition systems by handling pronunciation errors. A novel approach of forming phone confusion matrix for each dysarthric speaker using SSV from Phone-CAT model is discussed. Phone-CAT model handles the data efficiently by using less number of parameter for estimation. The SSV captures the context and phonetic information. It represent the enunciated phone using the dominant weight. This property is used to identify the mispronunciations made by each dysarthric speaker, by building speaker-specific Phone-CAT model. Using the phone confusion matrix, alternate pronunciations are formed in personalized speaker lexicons. These modified lexicons improves the performance of the dysarthric ASR system. This preliminary study shows that the proposed phone confusion matrix using SSV captures the speaker-specific pronunciation patterns and avoid the usage of decoded transcription. This approach has to be explored in detail to analyze, model the error pattern and handle the insertion and deletion errors which forms our future work.

9. References

[1] J. R. Duffy, “Motor speech disorders: clues to neurologic diagnosis,” in *Parkinsons Disease and Movement Disorders*, pp. 35–53, Springer, 2000.

[2] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, “Acoustic studies of dysarthric speech: Meth-

ods, progress, and potential," *Journal of communication disorders*, vol. 32, no. 3, pp. 141–186, 1999.

[3] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. E. Siong, "Severity-based adaptation with limited data for asr to aid dysarthric speakers," *PloS one*, vol. 9, no. 1, p. e86285, 2014.

[4] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[5] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pp. 255–256, ACM, 2007.

[6] E. Rosengren, P. Raghavendra, and S. Hunnicutt, "How does automatic speech recognition handle dysarthric speech?," *Lund Working Papers in Linguistics*, vol. 43, pp. 112–115, 2009.

[7] H. Kim, K. Martin, M. Hasegawa-Johnson, and A. Perlman, "Frequency of consonant articulation errors in dysarthric speech," *Clinical linguistics & phonetics*, vol. 24, no. 10, pp. 759–770, 2010.

[8] E. Sanders, M. B. Ruiter, L. Beijer, and H. Strik, "Automatic recognition of dutch dysarthric speech: a pilot study.," in *INTERSPEECH*, 2002.

[9] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proc. ICASSP*, pp. 4924–4927, IEEE, 2011.

[10] W. K. Seong, J. H. Park, and H. K. Kim, "Multiple pronunciation lexical modeling based on phoneme confusion matrix for dysarthric speech recognition," *Advanced Science and Technology Letters*, vol. 14, pp. 57–60, 2012.

[11] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 2, 2009.

[12] V. Manohar, C. Srinivas, S. Umesh, *et al.*, "Acoustic modeling using transform-based phone-cluster adaptive training," in *Proc. ASRU*, pp. 49–54, IEEE, 2013.

[13] C.-H. Wu, H.-Y. Su, and H.-P. Shen, "Articulation-disordered speech recognition using speaker-adaptive acoustic models and personalized articulation patterns," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 10, no. 2, p. 7, 2011.

[14] S.-O. Caballero-Morales and F. Trujillo-Romero, "Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition," *Expert Systems with Applications*, vol. 41, no. 3, pp. 841–852, 2014.

[15] P. Jyothi and E. Fosler-Lussier, "Discriminative language modeling using simulated asr errors.," in *INTERSPEECH*, pp. 1049–1052, 2010.

[16] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for oov keywords in the keyword search task," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pp. 416–421, IEEE, 2013.

[17] W. K. Seong, J. H. Park, and H. K. Kim, "Dysarthric speech recognition error correction using weighted finite state transducers based on context–dependent pronunciation variation," in *Computers Helping People with Special Needs*, pp. 475–482, Springer, 2012.

[18] W. K. Seong, J. H. Park, and H. K. Kim, "Performance improvement of dysarthric speech recognition using context-dependent pronunciation variation modeling based on kullback-leibler distance," *Advanced Science and Technology Letters*, vol. 14, no. 1, pp. 53–56, 2012.

[19] H. Christensen, P. D. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech.," in *INTERSPEECH*, pp. 1159–1163, 2013.

[20] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[21] B. Abraham, N. M. Joy, and N. K. Umesh, "A data-driven phoneme mapping technique using interpolation vectors of phone-cluster adaptive training," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pp. 36–41, IEEE, 2014.

[22] R. Sriranjani, S. Umesh, and M. Reddy, "Automatic severity assessment of dysarthria using state-specific vectors.," *Biomedical sciences instrumentation*, vol. 51, pp. 99–106, 2015.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU*, pp. 1–4, 2011.

[24] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Proc. ICSLP*, vol. 3, pp. 1962–1965, IEEE, 1996.

Determining an Optimal Set of Flesh Points on Tongue, Lips, and Jaw for Continuous Silent Speech Recognition

Jun Wang[1,2], *Seongjun Hahm*[1], *Ted Mau*[3]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]Department of Otolaryngology - Head and Neck Surgery
University of Texas Southwestern Medical Center, Dallas, Texas, United States
{wangjun, seongjun.hahm}@utdallas.edu; ted.mau@utsouthwestern.edu

Abstract

Articulatory data have gained increasing interest in speech recognition with or without acoustic data. Electromagnetic articulograph (EMA) is one of the affordable, currently used techniques for tracking the movement of flesh points on articulators (e.g., tongue) during speech. Determining an optimal set of sensors is important for optimizing the clinical applications of EMA data, due to the inconvenience of attaching sensors on tongue and other intraoral articulators, particularly for patients with neurological diseases. A recent study found an optimal set (tongue tip and body back, upper and lower lips) on tongue and lips for isolated phoneme, word, or short phrase classification from articulatory movement data. This four-sensor set, however, has not been verified in continuous silent speech recognition. In this paper, we investigated the use of data from sensor combinations in continuous speech recognition to verify the finding using a publicly available data set MOCHA-TIMIT. The long-standing speech recognition approach Gaussian mixture model (GMM)-hidden Markov model (HMM) and a recently available approach deep neural network (DNN)-HMM were used as the recognizers. Experimental results confirmed that the four-sensor set is optimal out of the full set of sensors on tongue, lips, and jaw. Adding upper incisor and/or velum data further improved the recognition performance slightly.

Index Terms: silent speech recognition, deep neural network, hidden Markov model, electromagnetic articulograph, articulation, dysarthria

1. Introduction

With the availability of affordable devices for tongue movement data collection, articulatory data have obtained interest not only in speech science [1, 2, 3, 4] but also in speech technology (i.e., automatic speech recognition) [5, 6]. First, articulatory data have been successfully used to improve the speech recognition accuracy [5]. Articulatory data are particularly useful when speech signals are noisy or low quality [7] for recognizing dysarthric speech [8, 9]. Second, when acoustic data is not available, a silent speech interface (SSI) based on articulatory data has potential clinical applications [10, 11]. An SSI recognizes speech from articulatory data only (without using audio data) [12, 13] and then drives a text-to-speech synthesizer for sound playback [14, 15]. For example, SSIs can be used to assist the oral communication for patients with severe voice disorders or without the ability to produce speech sounds (e.g., due to laryngectomy, a surgical removal of larynx due to treatment of laryngeal cancer) [16]. There are currently limited options to assist speech communication for those individuals (e.g., esophageal speech, tracheo-esophageal speech or tracheo-esophageal puncture (TEP) speech, and electrolarynx). These approaches, however, produce an abnormal sounding voice [17, 18], which impacts the quality of life of laryngectomees. Current text-to-speech technologies have been able to produce speech with natural sounding voice for SSIs [19]. One of the current challenges of SSI development is silent speech recognition algorithms (without using audio data) [10, 20] or mapping articulatory information to speech [21, 22, 23].

Electromagnetic motion tracking is one of the affordable, currently used technologies for tracking tongue movement during speech [19, 24, 25]. There are currently two commercially available devices, EMA AG series (by Carstens) and Wave system (by NDI, Inc.) [26]. Tongue tracking using electromagnetic devices is accomplished through attaching small sensors on the surface of tongue and other articulators. In prior work, the number of tongue sensors and their locations have been justified based on long-standing assumptions about tongue movement patterns in classic phonetics [27], or the specific purpose of the study. Other techniques that have been used to record non-audio articulatory information include ultrasound [28, 29], and surface electromyography (EMG) [30, 31].

Determining an optimal set of tongue sensors for speech production is significant for both science and technology. Scientifically, determining an optimal set of sensors can improve the understanding of the coordination of articulators for speech production [32]. Technologically, it can be helpful for clinical applications including (1) silent speech interfaces, (2) speech recognition with articulatory information [5, 33], and (3) speech training using real-time visual feedback of tongue movements [34, 35]. In literature, three or four EMA sensors on the tongue have been commonly used (e.g., [1, 3, 4, 5, 36, 37]). The use of more sensors than necessary comes at a cost for both researchers and subjects; the procedure for attaching sensors to the tongue is time intensive and can cause discomfort and therefore may limit the scope of EMA for practical use, particularly for persons with neurological diseases (e.g., Parkinson's disease [38] and amyotrophic lateral sclerosis [39]).

Here, *optimal* set means a sensor set that contains the least number of sensors that performs no worse than other sets with more sensors. There may be more than one optimal set with the same number of sensors.

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 79–85,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

Until recently, a study found two tongue sensors (Tongue Tip and Tongue Body Back) and two lip sensors (Upper Lip and Lower Lip) are optimal for isolated phoneme (vowels and consonants), word, and short phrase classification [32, 40]. The classification results based on data using the optimal set were not significantly different from these based on data from the full set with four tongue sensors (Tongue Tip, Tongue Blade, Tongue Body Front, and Tongue Body Back) plus the two lip sensors [32]. However, this set has not been verified in continuous silent speech recognition or speech recognition from both acoustic and articulatory data. If the two-tongue-sensor set can be confirmed for continuous speech recognition, it would be beneficial for future collection of a larger articulatory data set. Other studies compared the whole tongue and lips (e.g., [41] using ultrasound and optical data), but not on flesh points.

In this paper, we investigated the optimal set of tongue sensors for speaker-dependent continuous silent speech recognition (using articulatory data only) and speech recognition (using combined acoustic and articulatory data). The goals were (1) to confirm if more than two tongue sensors are unnecessary for continuous silent speech recognition and speech recognition using both acoustic and articulatory data when only tongue and lips are used, and (2) to provide a reference for choosing the number of sensors and their locations on the tongue, lips, jaw and other articulators for future studies. However, due to the space limitation, this paper did not verify if the hypothesized optimal four-sensor set is unique. The articulatory and acoustic data in the MOCHA-TIMIT data set [42] were used in this experiment. The MOCHA-TIMIT data set is appropriate for this study because it contains data collected from sensors attached on multiple articulators, including three sensors on the tongue, two on the lips, two on the incisors, and one on the velum. In addition, both MOCHA-TIMIT and the data set in [32] have tongue tip and body back (or dorsum). Thus the first goal of this paper became to verify if the tongue blade sensor is unnecessary in addition to the hypothesized optimal set [32, 40]. The traditional speech recognition approach Gaussian mixture model (GMM)-hidden Markov model (HMM) [5] and a recently available and promising approach deep neural network (DNN)-HMM [43, 44] were used.

2. Method

2.1. Data set

MOCHA (Multi-CHannel Articulatory)-TIMIT data set consists of simultaneous recordings of speech, articulatory movement, and other forms of data collected from 2 British English speakers (1 male - MSAK0 and 1 female - FSEW0) [42]. There are 920 sentences (extracted from TIMIT database) in total. Individual phonemes and silences within each sentence have been labeled.

The articulatory and acoustic data in MOCHA-TIMIT were collected using an Electromagnetic Articulograph (EMA, Carstens Medizinelektronik GmbH, Germany) by attaching sensors to upper lip (UL), lower lip (LL), upper incisor (UI), lower incisor (LI), tongue tip (TT), tongue blade (TB), tongue dorsum (TD), and velum (V) with 500 Hz sampling rate. Each sensor had x (front-back) and y (vertical) trajectories. Therefore, the acoustic data and the 16-dimensional x and y motion data obtained from UI, LI, V, UL, LL, TT, TB, and TD were used.

TT was 5-10 mm to the tongue apex; TB was 2-3 cm from TT; TD was 2-3 cm from TB [42]. This roughly matched with the tongue tip sensor in [32, 40], which was also 5-10 mm to tongue apex, and the tongue body back in [32, 40], which was about 40 mm from tongue tip. Thus, as mentioned earlier, the goal (1) in this paper became to verify if the middle tongue sensor (TB) was unnecessary.

2.2. Recognizers

A long-standing approach GMM-HMM and a promising approach DNN-HMM were used as the recognizers in this experiment.

2.2.1. Gaussian Mixture Model-Hidden Markov Model

GMM-HMM has been used in speech recognition for decades [45]. The core idea of GMM is compact representation of distribution using means and variances. GMM is a generative model and trained to represent as closely as possible the distribution (e.g., using means and variances) of training data. In many applications, the number of mixtures for GMMs is adjusted to avoid overfitting.

2.2.2. Deep Neural Network-Hidden Markov Model

DNN-HMM recently attracted the interests of speech recognition researchers because it showed a significant performance improvement compared with GMM-HMM when replacing GMM to DNN in (acoustic) speech recognition [44, 46]. We adopted the DNN training approach based on restricted Boltzmann machines (RBMs) [47].

The DNN (stacked RBMs) were subsequently fine-tuned using the backpropagation algorithm. A detailed explanation and discussion of the DNN can be found in [47, 48].

2.3. Experimental setup

Data from individual sensors or combinations of sensors were used in speech recognition experiments (from articulatory data only or from combined acoustic and articulatory data). The recognition performances obtained from individual sensors or their combinations were compared to determine (1) if Tongue Blade was unnecessary in addition to the other two tongue sensors and lips (Tongue Tip, Tongue Dorsum, Upper Lip, and Lower Lip), and (2) if the performance improved when more sensor's data (e.g., upper incisor and velum) were added.

In each experiment, a 5-fold cross validation strategy with a jackknife procedure was performed to set training and test sets in the experiment [42, 49]. In each of the five executions, a group of 92 sentences were selected for test with the remaining 368 sentences for training. Due to the high degree of variation in the articulation across speakers and there were only two speakers in MOCHA-TIMIT, speaker-dependent recognition was conducted. The average training data length for each cross validation became 21.3 mins (368 sentences) for the female speaker and 20.6 mins (368 sentences) for the male speaker. The average test data length along 5 cross validations was 5.3 mins (92 sentences) for the female speaker and 5.2 mins (92 sentences) for the male speaker, respectively.

Articulatory features were extracted from the corpus using EMAtools [50]. The original articulatory features and their first and second derivatives were concatenated to build various dimensional feature vectors for each set of sensors. The "breath" segments were merged with "silence" for both training and testing [49]. The input features in DNN were a concatenation of articulatory feature vectors (number of sensors × 2-dimension articulatory movement data + Δ + $\Delta\Delta$) with 9

Table 1: *Experimental setup.*

Articulatory Feature	
Low pass filtering	40 Hz cutoff, 5th order Butterworth
Sampling rate	100 Hz (downsampled from 500 Hz)
Feature vector	articulatory movement vector + Δ + $\Delta\Delta$ (e.g., 6 dim. for 1 sensor, 48 dim. for 8 sensors)
Acoustic Feature	
Sampling rate	16 kHz
Feature vector	MFCC vector (13 dim.) + Δ + $\Delta\Delta$ (39 dim.)
Frame size	25 ms
Common	
Frame rate	10 ms
Mean normalization	Applied
GMM-HMM topology	
Monophone	context-independent 137 states (44 phones $\times$ 3 states, 5 states for silence), $\approx$ 14 mixtures 3-state left to right HMM
Training method	Maximum likelihood estimation
DNN-HMM topology	
Monophone	context-independent input layer dimension varies based on the set of sensors (e.g., 54 for 1 sensor, 432 for 8 sensors) 137 output layer dimension (including 5 outputs for silence) 1,024 nodes for each hidden layer 1 to 6-depth hidden layers
Training method	RBM pre-training, back-propagation
Language model	bi-gram phoneme language model

frames (4 preceding, current, and 4 succeeding frames). As it concatenates multiple feature vectors in the time domain, DNN has time-dependent context information which HMM takes using multiple states [43, 51]. Mel-frequency cepstral coefficients (MFCCs) were extracted from the acoustic data and used as the acoustic features in the recognition experiments.

The GMM-HMM system was trained using maximum likelihood estimation (MLE) without using segment information provided in MOCHA-TIMIT corpus (flat initialization). The DNN-HMM system was pre-trained using contrastive-divergence algorithm on RBMs and fine-tuned using back-propagation algorithm. A bi-gram phoneme language model was trained using all 44 phonemes provided in label files of the corpus.

Table 1 lists the details of the experimental setup and major parameters in GMM-HMM and DNN-HMM. The training and decoding were performed using the Kaldi speech recognition toolkit [52].

A phoneme error rate (PER) was used as a performance measure, which is the ratio of the sum of the number of errors over the total number of phonemes. The PER is represented by

$$\text{PER} = (S + D + I)/N \tag{1}$$

where S represents the number of substitution errors, D is the number of deletion errors, I stands for the number of insertion errors, and N is the total number of phonemes in the test set. For DNN, we conducted experiments using 1 to 6 hidden layers and the best performance was reported. Finally, the PERs from each test group in the 5-fold cross validation were averaged as the overall PER.

3. Results and Discussion

Experimental results are shown in Figures 1 to 4 and discussed below. Figures 1 and 2 show the silent speech recognition performance on individual or combinations of sensors for both speakers using GMM-HMM or DNN-HMM, respectively. Figures 3 and 4 give the speech recognition from MFCCs plus individual or combinations of sensors' data using GMM-HMM and DNN-HMM, respectively.

3.1. General observations

First, the recognition performances obtained from individual sensor's data had consistently lower performance (higher PERs) than from the combinations of sensors (Figures 1 to 4). Although it seems intuitive, to our knowledge, this is the first time the individual EMA sensor's performance were examined in continuous silent speech recognition or speech recognition from combined acoustic and articulatory data.

Second, when the performances obtained using data from individual sensors were compared, upper incisor (UI) and velum (V) had the worst performance; the three individual tongue sensors had a similar performance and were the best among all sensors; lip sensors were between the tongue sensors (TT, TB, TD) and UI and velum (V). This finding is highly consistent with the descriptive knowledge in classic phonetics that tongue is the primary articulator [27].

3.2. {TT, TD, UL, LL} and other combinations

Silent speech recognition performance substantially degraded if any of the sensor in previously found optimal four-sensor set (i.e., TT, TD, UL, and LL, marked bold in Figures 1 and 2) was not used [32]. The optimal set of sensors using GMM-HMM and articulatory data yielded a PER of 42.0% and 40.9% for the female and male speakers, respectively. DNN-HMM with this optimal set yielded a PER of 38.2% and 36.5% for the female and male speakers, respectively.

As TB, UI, LI (jaw), or all of the three sensors' data were added on top of the four-sensor set, there was no improvement using GMM-HMM, but a slight improvement using DNN-HMM. When using all sensors' (including V) data together, a substantial improvement was obtained using either GMM-HMM or DNN-HMM.

These results suggest the four-sensor set ({TT, TD, UL, LL}) was an optimal set for silent speech recognition out of the full set of sensors on the tongue, lips, and jaw. However, adding extra data source (e.g., UI and V) could still improve the performance.

Speech recognition from combined acoustic and articulatory data (Figures 3 and 4) also substantially degraded if any of the sensor in {TT, TD, UL, and LL} was missing, for recognizers. However, GMM-HMM and DNN-HMM results showed different patterns when adding more sensors data to {TT, TD, UL, LL}. GMM-HMM showed no improvement to the optimal set (23.0% for female and 22.6% for male) when adding more sensor's data (22.7% for female and 22.8% for male); while DNN-HMM (19.7% for female and 19.5% for male) showed significant error reduction compared to the optimal set (20.4% for female and 20.3% for male). This observation suggests DNN has more potential than GMM to incorporate more data sources to further improve the recognition performance.

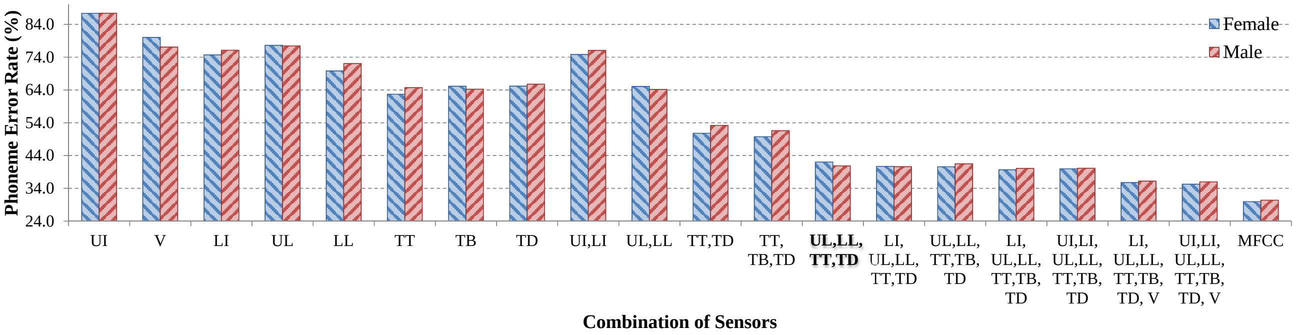


Figure 1: *Phoneme Error Rates (PER; %) obtained using GMM-HMM and articulatory features.*

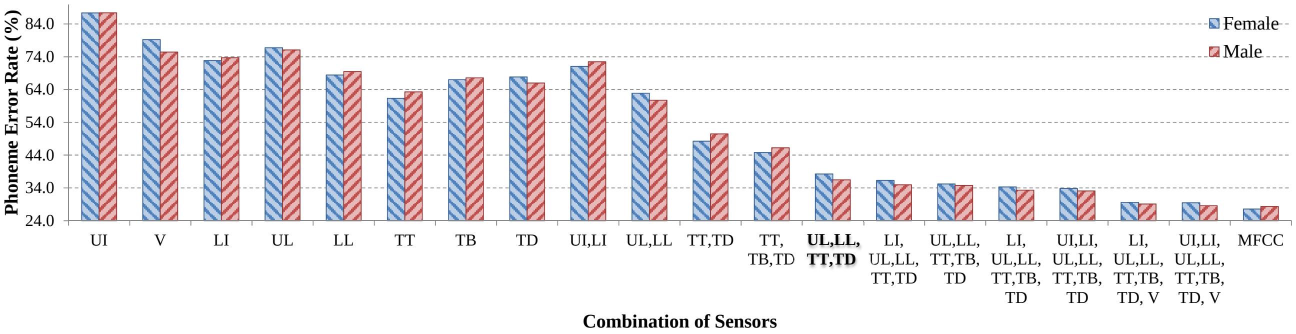


Figure 2: *Phoneme Error Rates (PER; %) obtained using DNN-HMM and articulatory features.*

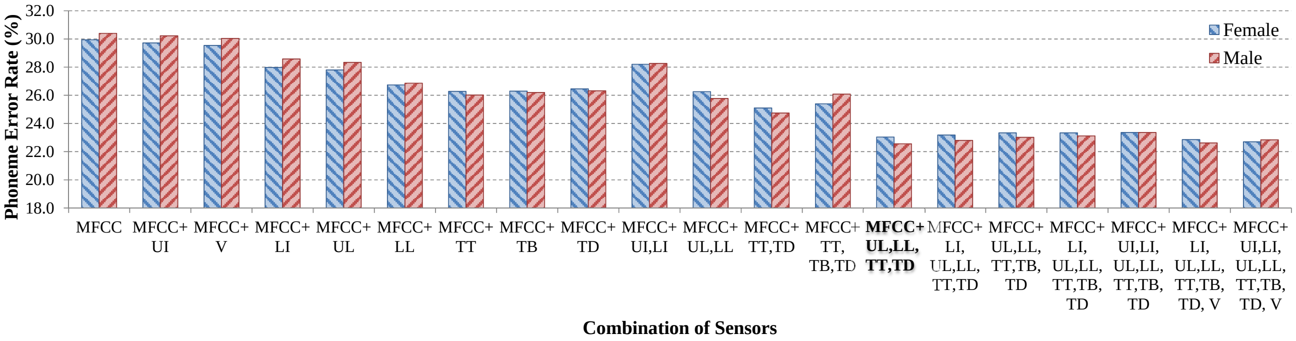


Figure 3: *Phoneme Error Rates (PER; %) obtained using GMM-HMM and combined articulatory and acoustic features.*

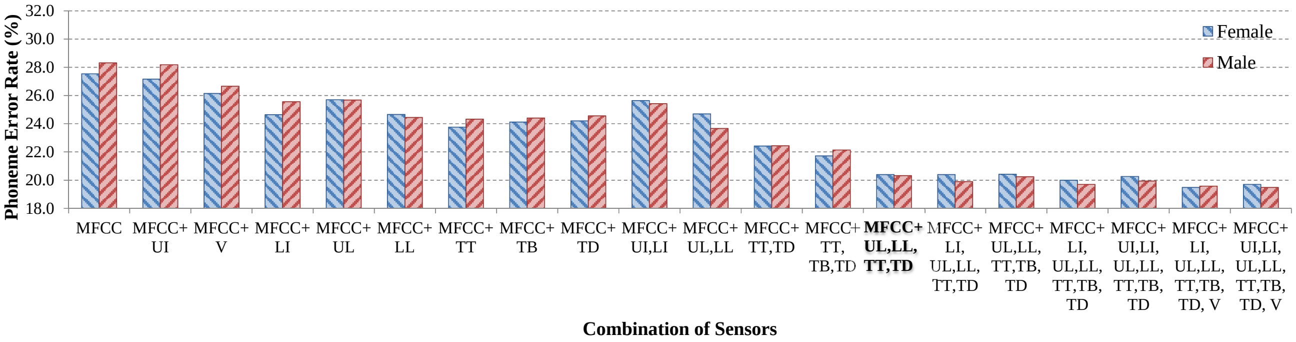


Figure 4: *Phoneme Error Rates (PER; %) obtained using DNN-HMM and combined articulatory and acoustic features.*

The most important conclusion from the results above may be, for future studies in which data are collected only from tongue, lips, or jaw (i.e. not from velum), {TT, TD, UL, LL} is an optimal set for silent speech recognition or speech recognition from combined acoustic and articulatory data. However, adding upper incisor and/or velum data can still further improve the performance slightly.

3.3. {TT, TD, UL, LL} vs {TT, TB, TD, UL, LL}

Table 2 lists the results obtained from {TT, TD, UL, LL} and {TT, TB, TD, UL, LL} to provide a close-up performance comparison of the two sets, which further confirms adding TB

Table 2: *Phoneme Error Rates (PER; %) obtained from sensor combination {TT, TD, UL, LL} and {TT, TB, TD, UL, LL}.*

			Combination of Sensors		Performance
Speaker	Model	Feature	TT,TD,UL,LL	TT,TB,TD,UL,LL	Difference
Female	GMM-HMM	EMA	42.04	40.60	+1.44
		MFCC + EMA	23.04	23.34	-0.30
	DNN-HMM	EMA	38.24	35.20	+3.04
		MFCC + EMA	20.40	20.42	-0.02
Male	GMM-HMM	EMA	40.88	41.48	-0.60
		MFCC + EMA	22.56	23.02	-0.46
	DNN-HMM	EMA	36.46	34.74	+1.72
		MFCC + EMA	20.32	20.24	+0.08
Average					+0.61

(Tongue Blade) did not significantly improve the speech recognition performance in addition to {TT, TD, UL, LL}. The right-most column of Table 2 lists the performance difference between {TT, TD, UL, LL} and {TT, TB, TD, UL, LL} (positive means a better performance with TB; negative means worse performance). The average performance difference of the two sensor sets in all eight configurations (female vs male speaker, GMM vs DNN, with or without MFCC) was +0.61, which means adding TB reduced only 0.61% of PER.

3.4. {TT, TD, UL, LL} may not be the only four-sensor optimal set

The four-sensor set ({TT, TD, UL, LL}) may be just one of the possible optimal four-sensor sets, because of the high coupling of adjacent parts [3]. Figures 1 to 4 also show the three tongue sensors, TT (Tongue Tip), TD (Tongue Dorsum) and TB (Tongue Blade) have no significant differences in performance when used individually, which may suggest they are interchangeable. In other words, any two tongue sensors may achieve no significant difference in recognition performance with {TT, TD}. A further analysis using data from all tongue sensor pairs is needed to test this hypothesis.

Nevertheless, we still suggest {TT, TD} as the optimal tongue sensor pair, since TT and TD are anatomically farther apart from each other than other tongue sensor pairs, thus TT and TD may be more independent and have less redundant information. In addition, from the user's (subject) perspective, the sensor location on the tongue may not matter, as long as they are in the comfortable zone (from tongue tip to tongue body back).

3.5. Velum sensor

Adding velum (V) data in addition to other sensors always improved the speech recognition performance, although velum in isolation achieved the worse performance. Velum is the primary articulator for controlling nasal sounds in English (e.g., /m/ and /n/). Velum provides unique information that other articulators do not. However, we still do not think attaching sensors on the velum is suitable for practical use of EMA, considering the trade-off of the discomfort of attaching velum sensor on subjects and the slight improvement of recognition performance.

3.6. DNN-HMM outperformed GMM-HMM

DNN-HMM outperformed GMM-HMM in all experimental configurations (Figures 1 to 4). Although the focus of this paper was not comparing GMM-HMM and DNN-HMM, the results indicate the DNN-HMM outperformed GMM-HMM in both silent speech recognition and speech recognition from combined acoustic and articulatory data. This finding is consistent with the recent literature in silent speech recognition [53], acoustic speech recognition [44, 48], and speech recognition from combined acoustic and articulatory data [46, 54]. We expect DNN-HMM has potential to further improve the recognition performance from articulatory data or from combined acoustic and articulatory data with a better structure or when combined with other approaches (e.g., speaker adaptation [55]).

4. Conclusions and Future Work

In this paper, we have confirmed a previously found optimal set of sensors on the tongue and lips (Tongue Tip, Tongue Dorsum, Upper Lip and Lower Lip) [32] through experiments with continuous silent speech recognition and speech recognition from combined acoustic and articulatory data, when only tongue, lips, upper incisor, and lower incisor data are available (i.e., no velum data). Although velum data can further (slightly) improve the recognition performance on top of the four-sensor set, it is not recommended for practical use because it causes discomfort for subjects. In addition, the four-sensor set may not be unique, since the individual tongue sensors have no significant accuracy difference. Finally, DNN-HMM outperformed GMM-HMM in both silent speech recognition and speech recognition from combined acoustic and articulatory data.

These findings provide a reference for future relevant studies on choosing the number of sensors and their locations on the tongue. However, as mentioned earlier, determining an appropriate set of sensors may depend on the specific purpose of the study. For example, a sensor on the side of the tongue may be used in studies that focus on lateral tongue curvature during speech production [56, 57].

Future work includes (1) verifying if TT, TB, and TD are interchangeable, or determining if {TT, TD, UL, LL} is the unique four-sensor optimal set, and (2) sensor combinations in speaker-independent silent speech recognition experiments [58, 59, 54].

5. Acknowledgment

This work was supported by the National Institutes of Health (NIH) through grants R03 DC013990 and R01 DC013547. We would like to thank Dr. Jordan R. Green, Dr. Ashok Samal, and the support from the Communication Technology Center, University of Texas at Dallas.

6. References

[1] J. S. Perkell and W. L. Nelson, "Variability in production of the vowels /i/ and /a/," *The Journal of the Acoustical Society of America*, vol. 77, no. 4, pp. 1889–1895, 1985.

[2] J. Westbury, "X-ray microbeam speech production database users handbook," *University of Wisconsin*, 1994.

[3] J. R. Green and Y.-T. Wang, "Tongue-surface movement patterns during speech and swallowing," *The Journal of the Acoustical Society of America*, vol. 113, no. 5, pp. 2820–2833, 2003.

[4] J. Wang, J. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[5] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[6] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magami-Doss, and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 jhu summer workshop," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, 2007, pp. IV–621–IV–624.

[7] K. Kirchhoff, G. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319, 2012.

[8] F. Rudzicz, "Using articulatory likelihoods in the recognition of dysarthric speech," *Speech Communication*, vol. 54, no. 3, pp. 430 – 444, 2012.

[9] ——, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2011.

[10] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[11] J. Wang, A. Samal, J. Green, and T. Carrell, "Vowel recognition from articulatory position time-series data," in *Proc. of International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2009, pp. 1–6.

[12] J. Wang, A. Balasubramanian, L. Mojica de la Vega, J. Green, A. Samal, and B. Prabhakaran, "Word recognition from continuous articulatory movement time-series data using symbolic representations," in *Proc. of Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Grenoble, France, 2013, pp. 119–127.

[13] J. Wang, A. Samal, J. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 4985–4988.

[14] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe, "Recent improvements on microsoft's trainable text-to-speech system-whistler," in *Proc. of ICASSP*, vol. 2, Munich, Germany, 1997, pp. 959–962.

[15] S. Manitsaris, B. Denby, F. Xavier, J. Cai, M. Stone, P. Roussel, and G.Dreyfus, "An open source speech synthesis module for a visual-speech recognition system," in *Proc. of Acoustics*, Nates, France, 2012, pp. 3937–3941.

[16] B. Bailey, J. Johnson, and S. Newlands, *Head & neck surgery–otolaryngology*. Lippincott Williams & Wilkins, 2006, vol. 1.

[17] H. Liu and M. Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332, 2007.

[18] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham, "Reconstructing the Voice of an Individual Following Laryngectomy," *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–66, 2011.

[19] J. Wang, A. Samal, and J. Green, "Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph," in *Proc. of ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, Baltimore, USA, 2014, pp. 38–45.

[20] J. Wang, "Silent speech recognition from articulatory motion," Ph.D. dissertation, The University of Nebraska-Lincoln, 2011.

[21] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech and Language*, 2015.

[22] M. W. Marlene Zahner, Matthias Janke and T. Schultz, "Conversion from facial myoelectric signals to speech: A unit selection approach," in *Proc. of INTERSPEECH*, 2013, pp. 1331–1335.

[23] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech and Language*, 2015.

[24] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical engineering & physics*, vol. 30, no. 4, pp. 419–425, 2008.

[25] R. Hofe, J. Bai, L. A. Cheah, S. R. Ell, J. M. Gilbert, R. K. Moore, and P. D. Green, "Performance of the MVOCA silent speech interface across multiple speakers," in *Proc. of INTERSPEECH*, 2013, pp. 1140–1143.

[26] J. Green, J. Wang, and D. L. Wilson, "Smash: A tool for articulatory data processing and analysis," in *Proc. of INTERSPEECH*, Vancouver, Canada, 2013, pp. 1331–1335.

[27] P. Ladefoged and K. Johnson, *A Course in Phonetics, 6th Edition*. Wadsworth Cengage Learning, 2011.

[28] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[29] B. Denby, J. Cai, P. Roussel, G. Dreyfus, L. Crevier-Buchman, C. Pillot-Loiseau, T. Hueber, G. Chollet *et al.*, "Tests of an interactive, phrasebook-style, post-laryngectomy voice-replacement system," in *Proc. of ICPhS XVII*, Hong Kong, 2011, pp. 572–575.

[30] C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Communication*, vol. 52, no. 4, pp. 354–366, 2010.

[31] Y. Deng, J. Heaton, and G. Meltzner, "Towards a practical silent speech recognition system," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 1164–1168.

[32] J. Wang, J. Green, and A. Samal, "Individual articulator's contribution to phoneme production," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7785–7789.

[33] F. Rudzicz, "Correcting errors in speech recognition with articulatory dynamics," in *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 60–68.

[34] W. Katz, T. F. Campbell, J. Wang, E. Farrar, C. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "Opti-speech: A real-time, 3d visual feedback system for speech training," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 1174–1178.

[35] P. Badin, A. B. Youssef, G. Bailly, F. Elisei, and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," in *Proc. of SLATE workshop*, 2010.

[36] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria." *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[37] F. H. Guenther, C. Y. Espy-wilson, S. E. Boyce, M. L. Matthies, M. Zandipour, J. S. Perkell, P. Frank, and H. Guenther, "Articulatory tradeoffs reduce acoustic variability during american english /r/ production," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2854–65, 1999.

[38] S. Hahm and J. Wang, "Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data," in *Proc. of INTERSPEECH*, 2015.

[39] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, "Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behavioral Neurology*, no. 183027, pp. 1–11, 2015.

[40] J. Wang, A. Samal, P. Rong, and J. R. Green, "An optimal set of flesh points on tongue and lips for speech movement classification," *Journal of Speech, Language, and Hearing Research*, In press.

[41] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Phone recognition from ultrasound and optical video sequences for a silent speech interface," in *Proc. of INTERSPEECH*, 2008, pp. 2032–2035.

[42] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. of ICSLP*, Beijing China, 2000, pp. 145–148.

[43] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping," in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 1297–1301.

[44] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at microsoft," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 8604–8608.

[45] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[46] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Cross-corpus and cross-linguistic evaluation of a speaker-dependent DNN-HMM ASR system using EMA data," in *Proc. of Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, 2013.

[47] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[48] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[49] E. Uraga and T. Hain, "Automatic speech recognition experiments with articulatory data," in *Proc. of INTERSPEECH*, Pittsburgh, USA, 2006, pp. 353–356.

[50] N. Nguyen, "A MATLAB toolbox for the analysis of articulatory data in the production of speech," *Behavior Research Methods, Instruments, & Computers*, vol. 32, no. 3, pp. 464–467, 2000.

[51] A.-R. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[52] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and V. K., "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, Waikoloa, USA, 2011, pp. 1–4.

[53] S. Hahm and J. Wang, "Silent speech recognition from articulatory movements using deep neural network," in *Proc. of the International Congress of Phonetic Sciences*, 2015.

[54] S. Hahm, D. Heitzman, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015.

[55] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7947–7951.

[56] J. Wang, W. Katz, and T. F. Campbell, "Contribution of tongue lateral to consonant production," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 174–178.

[57] A. Ji, J. Berry, and M. Johnson, "The electromagnetic articulography mandarin accented english (ema-mae) corpus of acoustic and 3d articulatory kinematic data," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 7719–7723.

[58] J. Wang, A. Samal, and J. Green, "Across-speaker articulatory normalization for speaker-independent silent speech recognition," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 1179–1183.

[59] J. Wang and S. Hahm, "Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training," in *Proc. of INTERSPEECH*, 2015.

Analysis of Dysarthric Speech using Distinctive Feature Recognition

Ka Ho Wong [1], *Yu Ting Yeung* [2], *Patrick C. M. Wong* [3],*Gina–Anne Levow* [4] *and H. Meng* [1,2]

[1] Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
[2] Stanley Ho Big Data Decision Analytics Research Centre,
[3] CUHK-Utrecht University Centre for Language, Mind and Brain,
[3] Department of Linguistics and Modern Languages,
The Chinese University of Hong Kong, Hong Kong SAR, China
[4] Department of Linguistics, University of Washington, Seattle, WA USA

khwong@se.cuhk.edu.hk, ytyeung@se.cuhk.edu.hk, p.wong@cuhk.edu.hk, levow@uw.edu,
hmmeng@se.cuhk.edu.hk

Abstract

Imprecise articulatory breakdown is one of the characteristics of dysarthric speech. This work attempts to develop a framework to automatically identify problematic articulatory patterns of dysarthric speakers in terms of distinctive features (DFs), which are effective for describing speech production. The identification of problematic articulatory patterns aims to assist speech therapists in developing intervention strategies. A multilayer perceptron (MLP) system is trained with non-dysarthric speech data for DF recognition. Agreement rates between the recognized DF values and the canonical values based on phonetic transcriptions are computed. For non-dysarthric speech, our system achieves an average agreement rate of 85.7%. The agreement rate of dysarthric speech declines, ranging between 1% to 3% in mild cases, 4% to 7% in moderate cases, and 7% to 12% in severe cases, when compared with non-dysarthric speech. We observe that the DF disagreement patterns are consistent with the analysis of a speech therapist.

Index Terms: speech recognition, distinctive feature, multilayer perceptron, dysarthric speech

1. Introduction

Dysarthria is a speech disorder caused by disturbances in the muscular control of the speech production mechanism [1]. Stroke, Parkinson's disease, cerebral palsy, amyotrophic lateral sclerosis and others nervous system-related diseases may cause dysarthria. Dysarthria affects millions of adults around the world, especially their effective speech communication in daily life. Speech-related problems include respiration, phonation, articulation and resonance. Symptoms that emerge in speech signals include hoarseness in voice quality, imprecise segmental articulation, excessive nasalization, as well as disordered prosody. All are detrimental to speech intelligibility.

Treatment of dysarthria involves perceptual assessment to characterize the problematic articulatory patterns, devise intervention strategies and monitor progress. Speech therapists generally listen carefully to dysarthric speech, possibly multiple times, in order to monitor progress, and such a process is costly. The situation calls for data-driven, computational techniques that analyze the problematic articulatory patterns of dysarthric speakers, in an attempt to assist human efforts in analysis to inform the development of intervention strategies.

Articulatory features describe the place and manner of articulation in speech production. They have been well-studied in the context of speech technology development, articulatory feature recognition with multiplayer perceptrons (MLPs) in telephone speech [2], and articulatory feature recognizer for dysarthric speech using neural networks and support vector machines [3] [4]. In particular, distinctive features (DFs) are a type of articulatory feature that also describe the general characteristics and acoustic consequences of the constrictions within the vocal tract [5]. DF have been shown to be well-identifiable from speech signals [5] [6], which motivates us to study the use of DFs in the analysis of dysarthric speech.

We aim to identify problematic articulatory patterns of dysarthric speech in terms of DFs. We apply an MLP-based DF recognition system on both dysarthric and non-dysarthric speech data from the TORGO corpus [7]. We compare the DF recognition results between dysarthric and non-dysarthric speech, with the DF reference derived from canonical pronunciations. For dysarthric subjects, we observe that the agreement rates of the DFs corresponding to poor articulation are significantly lower than those of the non-dysarthric subjects. We also note the relationships between the problematic articulatory patterns and the lower agreement rates of the corresponding DFs.

In the next section, we discuss the dysarthric corpus used for this study. In Section 3, we describe the development of a DF recognition system and the procedures to utilize the recognition results. In Section 4, we compare the results between manual analysis of the data based on Frenchay Dysarthric Assessment (FDA) [8] and the automatic DF recognition. We conclude our work in Section 5.

2. Dysarthric Speech

The TORGO (LDC2012S02) [7] corpus is a dysarthric speech corpus. The corpus includes 8 dysarthric subjects (3 females and 5 males) and 7 non-dysarthric subjects (4 male and 3 females). 7 dysarthric subjects are cerebral palsy and 1 is amyotrophic lateral sclerosis. There are 5 types of tasks in TORGO: recording articulatory movement tasks such as repeating "Ah-P-Eee", picture description, actions such as relaxing the mouth in its normal position, single word utterances such as saying

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 86–90,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

Dysarthric Subjects		Control Speakers	
Speaker ID	**Number of utterances**	**Speaker ID**	**Number of utterances**
F01	118	FC01	152
F03	545	FC02	965
F04	244	FC03	962
M01	371	MC01	726
M02	227	MC02	373
M03	406	MC03	799
M04	275	MC04	628
M05	332		

Table 1: *The number of utterances per speaker in the dataset.*

"yes" and sentential utterances such as "the quick brown fox jumps over the lazy dog". We focus on the single word tasks and sentence tasks. The dataset consists of 4,605 non-dysarthric speech utterances and 2,518 dysarthric speech utterances (Table 1). For the non-dysarthric speech, we further divide the data into a training set of 3,012 utterances and a testing set of 1,593 utterances. Both training and testing include male and female non-dysarthric subjects and no speakers overlap between training and testing.

3. Distinctive Feature Recognition

3.1. Phonetic-level Alignment of Speech Data

We perform automatic forced alignment on the TORGO speech data (both non-dysarthric and dysarthric) with the HTK toolkit [9]. We obtain phonetic-level alignments according to canonical pronunciations. We adopt the TIMIT phone set with modifications on the stops and diphthongs as in [2]. A stop like /p/ is split into a closure /pcl/ and release /p/. A diphthong is split into two phones. For example, /oy/ in "boy" is represented as the rounded portion /oy1/ followed by the unrounded portion /oy2/. We train an acoustic model based on the modified phone set with the TORGO non-dysarthric speech training dataset with the HTK scripts published in [10].

Phone deletion is observed in the dysarthric speech of the TORGO corpus as described in [11]. For example, M01 deletes /h/ in the word "house". We apply constrained grammars to handle phone deletions as shown in Figure 1. The constrained grammars are based on the phonetic-level canonical transcriptions, but an optional deletion path is provided for each phone. The current analysis is based on the "real" alignments which do not contain the deleted phones, although the statistics of phone deletion may be useful in future researches. An example of dysarthric speech alignment result is shown in Figure 2.

Transcription: /f iy/ ("fee")
Constrained grammar: [sil] [f] [sil] [iy] [sil]

Figure 1: *An example of a constrained grammar to handle phone deletion. The optional phones are braced by squared brackets* [].

Prompt: "The little schoolhouse stood empty" Aligned results:	
"The"	/dh ax/
"little"	/l ih tcl t/
"schoolhouse"	/_ kcl k uw l _ aw1 aw2 s/
"stood"	/_ tcl t uh dcl d/
"empty"	/eh m pcl p tcl t _/

Figure 2: *An aligned result for the M01's utterance. "_" represents missing phones. In [14], the authors reported M01 often omitted the initial /s/ and /h/and such cases are captured in the alignment in this work.*

Group	Distinctive Features	Meaning
Tongue	High, Low, Front, Back [6]	Place of tongue in vowel
	Lateral, Anterior [6]	The tongue part and shape used to produce sound
	Dental [16] , Alveolar [16], Retroflex [19], Velar [16]	The tip/blade of tongue will be placed different places to form a constriction.
Lips	Rounded , Labial [6]	The shape of lips
Soft Palate	Nasal [6]	The soft palate is lowered
Glottis	Aspirated [17]	The glottis stays open during the release
Vocal cords	Voiced [18]	There is periodic vibration of the vocal cords
Articulator-free	Tense [20]	Tense vowels are more intense, of longer duration and articulated with a greater deviation of the vocal cavity from its rest position then the lax vowels
	Delayed Release [20]	Slow release of stop closure
	Consonantal [6]	The absence or modification of constrictions in oral cavity
	Continuant [6]	Forming of complete closure
	Strident [6]	Any obstacle being placed in the airway downstream from the constriction
	Sonorant [6]	Pressure does not build up behind the constriction

Table 2: *The 21 DFs and their brief descriptions.*

3.2. Distinctive Features

Phonemes in languages can be represented in terms of a vector of distinctive features (DF) that capture their characteristics [6]. DFs include articulator-bound features like high, back, which relate to the tongue. DFs also include articulator-free features, such as tense, which correspond to the level of articulatory movement. We allow three possible values for each DF: positive ("+"), negative ("-") and "don't care" ("*"). "Positive" means that the articulatory movement that produces the phoneme fit the definition of the DF. For example, nasal is positive for /m/, which indicates that when /m/ is produced, the soft palate is lowered. "Negative" means that the articulatory movement and acoustic consequences described by the DF must not be observed when the phoneme is produced. For example, /b/ must be un-aspirated ("-"). Otherwise, it will become /p/ ("+" aspirated). "Don't care" means that the DF is not distinctive to the phone (e.g., high in /p/), or irrelevant (e.g. tense for /p/). We have chosen to apply 21 DFs in this work and their brief definitions are listed in Table 2.

DFs describe specific articulatory movements in speech production and their acoustic consequences. When DFs are applied for analysis of dysarthric speech, they should be able to help identify the problematic articulatory patterns that can inform the development of intervention strategies.

3.3. DF Recognition with Multilayer Perception

To train a DF recognition system, we start from the non-dysarthric speech data from the TIMIT training set. The

R L	+	-	*
+	▨	▨	▨
-	▨	▨	▨
*			X

(a) Three-class setting

R L	+	-
+	▨	▨
-	▨	▨
*		X

(b) Two-class setting

Figure 3: *An example of substitution -- /sh/ → /t/. "*" means "don't care". The shaded regions represent the outputs that we are interested. "L" and "R" mean labelled and recognized values respectively. "X" shows how the tense value being recognized in two settings. Since the tense value in /sh/ is "*", we don't care it being recognized as "*" (a) or "-" (b)*

TIMIT (LDC93S1) [12] corpus is a non-dysarthric speech corpus from a wide variety of speakers. The corpus provides us 6,300 non-dysarthric utterances for initial model training. It contains phonetic-level transcriptions with manually adjusted time alignment.

We train a frame-based MLP classifier for each of the 21 DFs [13]. Each MLP classifier consists of three hidden layers with 50 x 12 x 50 units in the hidden layers and sigmoid activation based on the previous work [14]. For the input layer, each input feature vector consists of features from 9 consecutive frames centered on the frame of interest to include the left-right context [2]. For each frame, the feature is 39-dimenensional Mel-frequency cepstral coefficients (MFCC) (12 coefficients + log-energy + Δ + ΔΔ). The feature is normalized as zero mean and unit variance.

At the output layer, there are two possible configurations, either (a) with three-class "+", "-" or "don't care", or (b) with two-class "+" or "-". The different configurations have different confusion matrices (Figure 3). We choose the two-class configuration (b) as in Figure 3. The DF recognition problem is generally a binary decision problem as to whether the recognized value matches with the reference value. For a case labeled "don't care", it is irrelevant whether the classifier's output is "+" or "-", because the DF value does not affect the phone's identity. During the training of each DF, we skip the frames which are silent or labeled as "don't care", but we still include them into the feature vectors. The label with maximum posterior probability will be assigned to the frame [12].

We further adapt the TIMIT MLP classifiers with non-dysarthric speech data of the TORGO corpus. The initial weights of the adapted classifiers are the same as the weights in the TIMIT MLP classifiers. The weights are updated with the same training process.

Dysarthric Subjects		
Subjects ID	**Severity**	**Average Agreement Rate Difference of Individual DF**
F01	Severe	7.9%
M01	Severe	11.2%
M02	Severe	9.1%
M04	Severe	8.7%
M05	Moderate-to-severe	7.4%
F03	Moderate	4.1%
F04	Mild	1.2%
M03	Mild	2.8%

Table 3: *A comparison of severity and the average DF agreement rate degradation of individual subjects.*

During DF recognition, we apply all 21 DF classifiers on both dysarthric and non-dysarthric speech data to obtain the corresponding DF values ("+" or "-") at each frame. For the TIMIT corpus, we compare the recognized DF results with real transcriptions included in the corpus. For the TORGO corpus, we compare the results with the canonical DF transcriptions by assuming that the subjects intend to read the prompts correctly. This is appropriate for a real application where real transcriptions are not available immediately. We thus interpret the recognized results as the agreement rate between the recognition system and the canonical DF transcriptions. In computing the agreement rate of each DF, we only consider the frame situated at the middle of the start time and end time of a phone.

Figure 4 shows the performance of each DF on the TIMIT testing set with the TIMIT MLP recognition system. An average agreement rate of 91.9% suggests that the DF recognizer is well-trained with non-dysarthric speech, as compared with 92% average frame on phonological binary features achieved by [15]. Figure 4 also shows the performance of the adapted DF recognition system on the TORGO dysarthric and non-dysarthric speech data. On non-dysarthric speech of the TORGO corpus, the average agreement rate drops to about 85.7%. The slightly lower DF agreement rate of TORGO non-dysarthric speech is probably due to occasional pronunciation variation from canonical pronunciations.

The severity of each dysarthric subject is reported in [11]. The average reduction in agreement rates of each dysarthric subject is calculated by equation (1)

$$D_i = \frac{1}{N}\sum_{j=1}^{N}(T_j - A_{i,j}) \qquad (1)$$

where D_i is the average agreement rate reduction of dysarthric subject i, N is the total number of DFs, T_j is the average

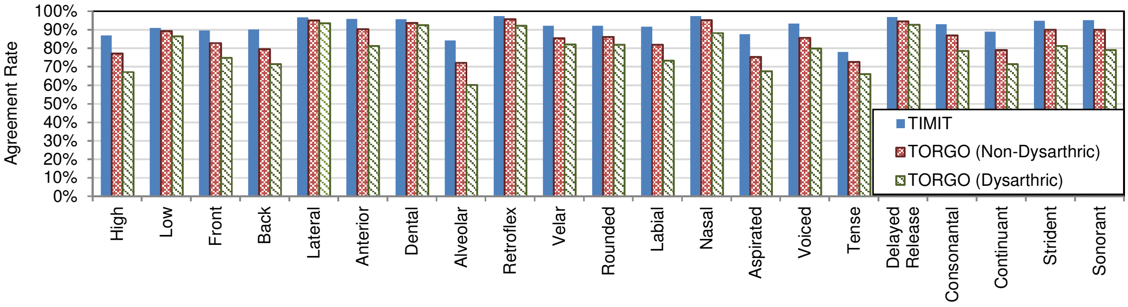
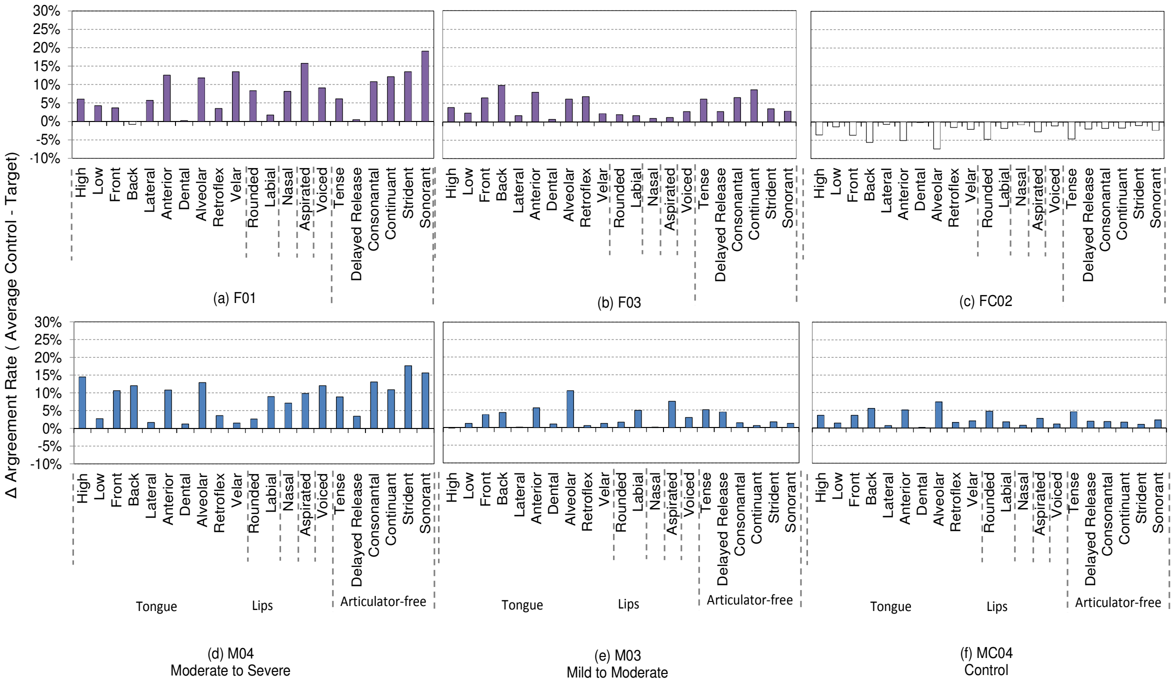


Figure 4: *The agreement rate of each DF between recognized results and canonical DFs.*



(a) F01

(b) F03

(c) FC02

(d) M04
Moderate to Severe

(e) M03
Mild to Moderate

(f) MC04
Control

Figure 5: *The difference between the average DF agreement rate from the control subjects and the corresponding DF agreement rate of each dysarthric subject. The agreement rates of most of the DF drop substainally for severely dysarthric subjects. The agreement rates in moderately and mildly dysarthric subjects only dropped in a few DFs.*

agreement rate of DF_j from all non-dysarthric subjects in TORGO shown in Figure 4, $A_{i,j}$ is the agreement rate of DF_j of dysarthric subject i.

The average reduction in DF agreement rates, D_i, is shown in Table 3. More severely dysarthric subjects have larger agreement rate reduction.

4. Discussion on Dysarthric Speech

4.1. Manual Analysis

A speech therapist has evaluated the severity of the dysarthric subjects in the TORGO corpus with Frenchay Dysarthric Assessment (FDA) [8]. FDA is one of the standard dysarthric speech assessments and includes 28 tests for different articulations. Each test is rated from “no abnormality” to “severe”. For speech production, there are tests of respiration, lips, jaw, palate, laryngeal production and tongue. There are also speech intelligibility tests at word, sentence and conversational levels. The FDA results provide us the reference to the severity of the dysarthric subjects on different articulatory dimensions.

We validate the recognized DF error patterns to the FDA results and the manual analysis from [11]. In [11], the authors studied 25% of the speech data of each dysarthric subject and identified the pronunciation error patterns of the individual subjects.

4.2. Severely Dysarthric Subjects

Figure 5 shows the drop in DF agreement rates for two severely dysarthric subjects (F01 and M04), one moderately dysarthric subject (F03), one mildly dysarthric subject (M03) and two non-dysarthric subjects (FC02 and MC04) for comparison to illustrate the relationship among the error patterns and agreement rates. FC02’s pronunciation is slightly better than that of MC04.

For the tongue-related DFs, F01 exhibits substantial drops in agreement rates on *anterior, alveolar* and *velar*. M04 also exhibits drops in agreement rates on *high, front, back, anterior* and *alveolar* relative to mildly dysarthric subjects. For F01 and M04, the speech therapist rated the correctness of articulation points and laboriousness of tongue motion as moderate-to-severe. This result is consistent with the reduction of tongue-related DFs agreement rates.

F01 and M04 also exhibit drops in agreement rates on *rounded* and *labial* respectively. Both of them are diagnosed with consistently poor lip movements by the speech therapist. Both of them have relatively poor DF agreement rates on *nasal* compared to mild subjects. The speech therapist also remarked that F01 has nasal emission problems. Although the DF results show M04 also has difficulty with *nasal*, the speech therapist reported that M04 only had slight problems with soft palate movement. Further analysis is necessary.

The DF results on *voiced* suggest that F01 and M04 may have problems in laryngeal production. In [11], the authors observed that the two subjects voice voiceless target consonants (prevocalic voicing problems). This observation agrees with the speech therapist’s findings that their voice production is inappropriate and ineffective in most situations.

For articulator-free DFs, the dysarthric subjects generally exhibit lower agreement rates on *consonantal*, *continuant* and *strident*. The trend is consistent with other consonant-related DFs. *Continuant* relates to the production of /f/ (“+”, no complete closure) and /p/ ("-", complete closure). The drop in *continuant* agreement rates of F01, M04 and F03 are higher than M03. The analysis in [11] also found that some fricatives (e.g. /f/) are replaced with stops (e.g. /p/) by F01 and F03 but not by M04. *Strident* affects fricatives such /f/ and /s/. In [11], the authors observed that F01 and M04 replace fricatives such as /f/, /s/ with non-fricatives such as /p/, /t/. We also observe the large agreement rate reductions on *strident* for F01 and M04.

There are substantial agreement rate reductions of *sonorant* for F01 and M04 (19.0% and 15.7% respectively). The results show that the subjects may have difficulty in building up pressure behind the constriction, which may be related to the lips problems described before.

Not all DFs exhibit these drops in agreement. The agreement rates on dental are similar among different dysarthric subjects. Some DFs may not be as useful in indicating the severity of the subjects. This is an area for future investigation.

4.3. Mildly and Moderately Dysarthric Subjects

The mildly dysarthric and moderately dysarthric subjects (M03 and F03) only exhibit slight agreement rate reductions for most DFs. In terms of DF results, the average agreement rates of F03 are lower than M03. The observation agrees with [11] that F03 is moderately dysarthric and M03 is mildly dysarthric. For F03, the agreement rates of tongue related DFs are worse than other articulator-bound DFs. The speech therapist also found that F03 had mild tongue-related problems.

5. Conclusions and Future Work

We compared the recognized DFs on dysarthric speech to prior results of manual analysis on the same dysarthric speech corpus. The general trends of reduced agreement are consistent with the analysis of the speech therapist and the observations of [11]. This indicates a potential way to automate analysis of dysarthric speech to assistant speech therapists for the development of intervention strategies. In the future, we plan to extend this framework to other languages such as Chinese. We will continue to improve the DF recognition system.

6. Acknowledgements

This project is partially sponsored by a grant from the Hong Kong SAR Government General Research Fund (reference no. GRF415513).

7. References

[1] D. B. Freed, *Motor Speech Disorders: Diagnosis & Treatment*. Clifton Park, NY: Delmar, Cengage Learning, 2012.

[2] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and O. Cetin, "Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech," in *Interspeech*, 2007.

[3] C. Middag, Bocklet T., Martens J.-P., and Nöth E., "Combining Phonological and Acoustic ASR-free Features for Pathological Speech Intelligibility Assessment," in *Interspeech*, Florence, Italy, 2011.

[4] F. Rudzicz, "Phonological Features in Discriminative Classification of Dysarthric Speech," in *International Conference on Acoustic, Speech and Signal Processing*, 2009.

[5] K. N. Stevens, "Toward a Model for Lexical Access Based on Acoustic Landmarks and Distinctive Features," *The Journal of the Acoustical Society of America*, vol. 111(4), pp. 1872-91, April 2002.

[6] K. N. Stevens, *Acoustic Phonetic*. Cambridge, MA: MIT Press, 1998.

[7] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO Database of Acoustic and Articulatory Speech from Speakers with Dysarthric Patient," *Language Resources and Evaluation*, vol. 46(4), pp. 523-541, 2012.

[8] P. M. Enderby, *Frenchay Dysarthria Assessment*. San Diego: College Hill Press, 1983.

[9] S. Young, J. Odell, D. Ollason, V. Valthcey, and P. Woodland, *The HTK Book*.: Cambridge University, 1995.

[10] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-Aligner: A Tool for Forced Alignment of Laboratory Speech," *Canadian Acoustics*, vol. 39.3, pp. 1192-193, 2011.

[11] K. Mengistu and F. Rudzicz, "Adapting Acoustic and Lexical Models to Dysarthric Speech," in *International Conference of Acoustic, Speech and Signal Processing*, 2011.

[12] J. S. Garofolo et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Lingusistic Data Consortium*, 1993.

[13] D. Johnson et al. (2004) ICSI QuickNet Software Package. [Online]. http://www1.icsi.berkeley.edu/Speech/qn.html

[14] P. K. Muthukumar and A. W. Black, "Automatic Discovery of a Phonetic Inventory for Unwritten Languages for Statistical Speech Synthesis," in *Internationl Conference of Acoustic, Speech and Signal Processing*, 2014.

[15] S. King and P. Taylor, "Detection of Phonological Features in Continuous Speech using Neural Networks," *Computer Speech and Language*, vol. 14(4), pp. 333-353, 2000.

[16] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Boston, MA: Wadsworth, Cengage Learning, 2009.

[17] R. Mannel. (2014, December) Phonetics and Phonlogy: Distinctive Features. [Online]. http://clas.mq.edu.au/speech/phonetics/phonology/features

[18] M. Halle and G. N. Clements, *Problem Book in Phonology: A Workbook for Introductory Courses in Linguistics and in Modern Phonology*. Cambridge, MA: MIT Press, 1983.

[19] S. R. Hamann, "The Phonetics and Phonology of Retroflexes," University of Utrecht, The Netherlands, PhD Dissertion 2003.

[20] N. Chomsky and M. Halle, *The Sound Pattern of English*. NY: Harper & Row, 1968.

[21] C. Middag, F. Hilgers, J.-P. Martens, M van den Brekel and R. van Son R. Clapham, "Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer," *Speech Communication*, vol. 59, pp. 44-54, January 2014.

Generating acceptable Arabic Core Vocabularies and Symbols for AAC users

E.A. Draffan, Mike Wald, Nawar Halabi, Ouadie Sabia[1]*, Wajdi Zaghouani* [2]

Amatullah Kadous, Amal Idris,[3] *Nadine Zeinoun, David Banes, Dana Lawand*[4]

[1]University of Southampton, UK
[2]Carnegie Mellon University, Qatar
[3]Hamad Medical Corporation, Qatar
[4]Mada Assistive Technology Center, Qatar

ead@ecs.soton.ac.uk, mw@ecs.soton.ac.uk, nh1g12@ecs.soton.ac.uk, o.sabia@soton.ac.uk, wajdiz@cmu.edu, tullahk@hotmail.com, aahmad2@hamad.qa, nzeinoun@mada.org.qa, dbanes@mada.org.qa, dlawand@mada.org.qa

Abstract

This paper discusses the development of an Arabic Symbol Dictionary for Augmentative and Alternative Communication (AAC) users, their families, carers, therapists and teachers as well as those who may benefit from the use of symbols to enhance literacy skills. With a requirement for a bi-lingual dictionary, a vocabulary list analyzer has been developed to evaluate similarities and differences in word frequencies from a range of word lists in order to collect suitable AAC lexical entries. An online bespoke symbol management has been created to hold the lexical entries alongside specifically designed symbols which are then accepted via a voting system using a series of criteria. Results to date have highlighted how successful these systems can be when encouraging participation along with the need for further research into the development of personalised context sensitive core vocabularies.

Index Terms: symbols, Augmentative and Alternative Communication, AAC, core vocabularies

1. Introduction

In the last few years it has become clear that many therapists and teachers working with individuals who have speech and language difficulties in the Arabic speaking Gulf area, are depending on westernized symbols and English core vocabularies. Issues around limited Arabic language knowledge and dependency on translations or working in English can cause difficulties for those who need Augmentative and Alternative forms of Communication (AAC) due to disabilities. Huer [1] reports that “observations of communication across cultures reveal that non-symbolic as well as symbolic forms of communication are culturally dependent” and her later work “suggests that consumers, families, and clinicians from some cultural backgrounds may not perceive symbols in the same way as they are perceived within the dominant European-American culture” [2].

With this in mind the Arabic Symbol Dictionary research team were determined to take a participatory approach to their project, involving AAC users and those supporting them as well as other researchers working in the field of Arabic linguistics and graphic design.

2. Background

Much has been written by speech and language therapists about the necessity for core vocabularies that have been adapted to suit symbol users who need to enhance their language skills [3], [4], [5] and [6]. Research has shown that with a few hundred of the most frequently used words 80% of one’s communication needs can be accommodated [7]. More recently concept coding [8] with the idea of mapping different symbol vocabularies along with a focus on psychosocial and environmental factors [9] to improve outcomes have been added to the mix. However, there is very little research that has been undertaken to provide therapists with suitable vocabularies for Arabic AAC users [10]. In English these vocabularies tend to be lists of frequently used words from spoken and written language across all age groups and some from AAC users. Despite considerable searching there are very few of these vocabularies available in Arabic with most coming from language learning or frequently used word lists with no specified ages or Arabic AAC users.

In some areas there is also a lack of understanding regarding the complexities of Arabic spoken and written language that disproportionately affect those who may have communication and reading difficulties [11], [12] and [13]. Usziel-Karl et al [13] cite several researchers in the course of their study concerning Arabic and Hebrew linguistic frameworks and discuss the “critical importance of morphology as the main organizing principle both of the lexicon and of numerous grammatical inflections”. The authors go on to point out the diglossia [two variations of a language in different social situations] nature of Arabic which means there is a ‘phonological distance [in grapheme-to-phoneme mapping] that has a negative impact on the acquisition of basic literacy skills in young Arabic children…” Words or word phrases (referents) may also be presented above or below a corresponding symbol, with changing forms depending on

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 91–96,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

grammatical status, gender and/or number plus many letters will change their shape depending on their position within a word.

The authors of this research and others have also found there are key cultural and family values/orientations that should be considered in order to increase the effectiveness of symbol-referent vocabulary interventions [14] with individuals who use AAC within Arab communities. To this end not only has research concentrated on word frequency lists and collating an AAC user core vocabulary, but also instigating a voting system for symbol acceptance, so that words or multiword/word phrases are represented by symbols that are suitable culturally, linguistically and for the settings in which they will be used.

3. Methodology for Building a Core Vocabulary

The building of an Arabic AAC core vocabulary is ongoing, but began with the collection of word lists used by AAC users, their families, carers, speech and language therapists and teachers in Doha (Qatar) (List a). Sixty three of these individuals joined an AAC forum and these participants have continued to work with the team as symbols for the vocabularies have been developed.

The initial aim was to collect around 100 localised Arabic most frequently used words and multiwords to compare with those already in use that were in English or translated into Arabic based on English core vocabularies. Participating therapists felt a further 400 words/multiwords would be the maximum the majority of their users would have in their communication books or devices. Most English speaking three year olds use over a thousand words [15] so it was essential that the fringe vocabulary should be enlarged with words specific to the environment and personal needs including Qatari colloquial words and place names as well as to be relevant to all ages.

Surveys of core vocabularies in Arabic have revealed that few are freely available [16] and even less make good companions when thinking of basic language and literacy learning for AAC users. In order to expand the list of 500 words a comparison was carried out against five other Arabic word frequency lists. Those for general conversation included the Kelly Project [17], 101languages.net 1000 most common spoken Arabic words and Aljazeera comments often using colloquial language [18]. The Supreme Education Council (SEC) literacy lists Grade 1,2,3 and Lebanese reading lists [19] have been used for literacy skill building in Modern Standard Arabic (MSA).

3.1. Building a vocabulary list analyser

An automatic system was developed that took as an input two main pieces of information:

List a: The list to be analyzed as a basis for the new core vocabulary list: This list could optionally have frequency of each entry included. If no frequency is available then a default value should be added to all the entries before running the program. Frequency in this case equated to how often a word was used. This frequency does not have to correspond to an actual frequency of occurrence in a text somewhere.

Lists b: Lists combining existing vocabularies from a number of sources with the same structure as List a. Multiple vocabularies are used in Lists b in an attempt to weight the occurrence of individual words. These vocabularies are ideally from different sources and should be large enough so that the frequencies of the entries listed are reliable.

The system produced three lists shown in Figure 1:

List 1: Initial list containing the words in List a (the in-put list to be analyzed) that did not occur in any of Lists b. This output only contained the words with no frequency scores.

List 2: The coverage list: containing the words that occurred in List a and at least once in a source vocabulary in Lists b. This output also contained scores for each word by source vocabulary list (each word was given several scores, one for each list in Lists b). Each score equals the frequency with which each word appeared in the list from Lists b, normalized by dividing the frequencies of each word by the sum of all frequencies in that list. The score was set to 0 if the word did not occur in that list.

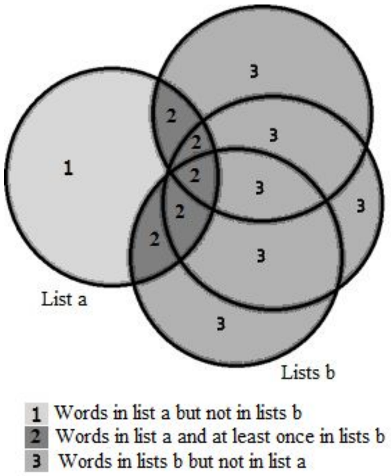
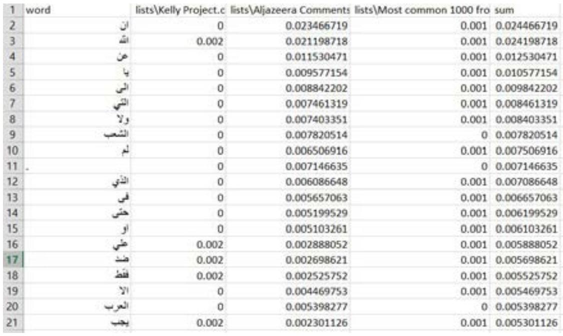

3
2
2
3
1
2
3
2
3
2
3
List a
3
Lists b
1 Words in list a but not in lists b
2 Words in list a and at least once in lists b
3 Words in lists b but not in list a

Figure 1. *Input lists (list a and lists b)*

List 3: Remaining word list: This list contained all the words that were in Lists b but were not contained in List a. This output also contained the scores for each word and is the example of the system in use (Figure 2). This is the list on which the comparison in the section 3.2 is based.



1	word	lists\Kelly Project.c	lists\Aljazeera Comments	lists\Most common 1000 fro	sum
2	ان	0	0.023466719	0.001	0.024466719
3	الله	0.002	0.021198718	0.001	0.024198718
4	عن	0	0.011530471	0.001	0.012530471
5	يا	0	0.009577154	0.001	0.010577154
6	الى	0	0.008842202	0.001	0.009842202
7	التي	0	0.007461319	0.001	0.008461319
8	ولا	0	0.007403351	0.001	0.008403351
9	الشعب	0	0.007820514	0	0.007820514
10	لم	0	0.006506916	0.001	0.007506916
11	.	0	0.007146635	0	0.007146635
12	الذي	0	0.006086648	0.001	0.007086648
13	في	0	0.005657063	0.001	0.006657063
14	حتى	0	0.005199529	0.001	0.006199529
15	او	0	0.005103261	0.001	0.006103261
16	علي	0.002	0.002888052	0.001	0.005888052
17	ضد	0.002	0.002698621	0.001	0.005698621
18	فقط	0.002	0.002525752	0.001	0.005525752
19	الا	0	0.004469753	0.001	0.005469753
20	العرب	0	0.005398277	0	0.005398277
21	يجب	0.002	0.002301126	0.001	0.005301126

Figure 2. *Example Output from lists viewed in Excel*

Figure 2. shows frequencies are normalized to allow source vocabularies to be compared (column one), this process can be problematic if the list is too small as the numbers may become too high and significantly affect results. Even if there is sufficient data, it is still imperative that an expert goes through the different output list to inspect the results, correct errors and choose the set of words to be added or removed from the input list. The scores given only act as a guide to assist the expert in the process.

In practical terms words with high scores in List 3 could be deemed suitable for inclusion in the Arabic Symbol Dictionary and added to List a. The system has been run repeatedly as lists have been added so that results become more robust.

3.2. Results of the Core Vocabulary building

When comparing the list provided by participants as examples of AAC users' vocabularies (List a), there were very small overlaps with those words most frequently found where the top words were based on very high frequency scores for those most commonly used (Lists b).

To provide an instant comparison between Output 1 and 3 the top 20 words translated from Arabic are listed below.

Output from 1 (List a) ordered by those most often used in AAC lists.

> *"I/me (am), go, ball, car, banana, on/to, thing/something, to, chair, clock/watch, want, in, sit, was, eat, bike, flower/rose, play, cup, door"*

Output from 3 (Lists b) ordered by frequency

> *"the, God, about, oh, to, which (masculine), and not, people, no, which (feminine), in, even, or, on, against, only, however, Arabs, must, order"*

Further analysis of the Lists b that were about spoken and colloquial language shows that nouns only made up 5% of the total list from the Kelly project, 25 to 30 % of the Aljazeera and Oweini-Hazoury lists, but 50% of the AAC lists. A concrete noun, even if it is considered part of a fringe vocabulary, is a much easier concept to illustrate with a symbol and may be seen as one of the early building blocks to language acquisition. Verbs, however are more complex and have low frequency rates; between 5 to 20 %. The Aljazeera list has the lowest and the AAC lists have the highest. The other parts of speech, equally pertinent in communication, such as adjectives, adverbs, prepositions, pronouns and conjunctions were found to be variably frequent from one list to another. The Aljazeera list has a quarter of its frequencies made up of prepositions, whereas Kelly's list, SEC and the AAC user list have only 5%. Conjunctions also show low frequencies through the lists in question; between 1% and 15%. It is worth mentioning that pronouns are totally nonexistent in Kelly's project list, either under their detached form or attached form. It should also be noted that therapists may choose nouns rather than pronouns for the purpose of symbol transparency. The other lists had less than 20% of pronouns all types combined. Arabic pronouns, and also some prepositions combine with nouns or with other parts of speech as single words, this morphological aspect could be the reason why their frequencies are rather undermined. Adverbs are also rarely listed, The Owein-Hazoury list has none; the highest adverb frequency is found in the 1000 most common Arabic words list (4%). In Arabic most adverbs of time and space are prepositional groups; typically a structure made of a preposition followed by a noun. This structural definition of adverbs explains the low number or even the lack of adverbs in some of the core vocabulary lists. The users would frame appropriate phrases to express adverbs by using existing prepositions combined with nouns.

Further confirmation for these differences in the frequency of various parts of speech was sought for the literacy skill vocabularies. The conversational based lists were replaced with reading lists forming Lists b. Arabic lists such as those used SEC and Arabic sight words [19]. It was found that in their top 100 frequently used words 30 and 38 were nouns respectively.

3.3. Discussion about the core vocabulary data collection

As can be seen from the top 20 words in List a and Lists b, both show nouns that would not be found in the top twenty frequently used words in an English core vocabulary and in reality would be considered fringe words. However, the lists do illustrate that in Arabic there are elements of the grammar that are equally as important such as conjunctions and prepositions.

There are considerable issues with the fact that root words in Arabic clearly appear within other words and this can affect the results as well as the fact that the lists collected from AAC users are based on popular use, rather than large scale frequency levels within a huge corpus. There will always be the need to improve outcomes by collecting more lists from AAC users in the future to improve the balance between words used for symbol communication and those based on frequency of use, although the latter informs vocabulary development

By using this system the combined AAC word lists from the Doha schools and clinics making up 'List a' once translated into English, could be compared to the Prenke Romich 100 Frequently Used Core Words [20], [21] (as Lists b). It was noted that the Doha Arabic AAC user list (List a) contained 38 nouns in the top 100 words compared to none appearing in the English core vocabulary. It has been said that in English the use of nouns goes from 7% in the top 100 words to 20% in the top 300 [22] whereas in MSA the corresponding frequency levels are 26% and 45% according to one of the largest frequency lists [23].

These results highlight the need for further exploration into this aspect of vocabulary building. In particular there is a need to collect more wide ranging conversations to evaluate the differences in the type of words and multiwords required to successfully build Arabic AAC personalised and context sensitive vocabularies. There is also the need to be aware of the differences in lists used for enhancing reading skills where MSA is used rather than the colloquial dialects of the area. A further distinction may be needed between adult and children's vocabularies where religious and social language requirements may impact on AAC use. The Speech and Language therapists attending meetings with the team also noted the importance of vocabularies sensitive to user's characters, interests and social setting commenting on dress and gender issues as well as being aware of the issues of using lists from AAC users of school age due to the lack of available adult AAC users in the region at the time of writing.

4. Methodology for Symbol Management

Just as it was found that there was a paucity of core AAC vocabulary lists in Arabic, the same could be said about the symbols provided for AAC devices. Some centres in Doha were providing specifically designed symbols for the Arabic culture, environment, social and personalised linguistic needs but there were no adapted symbol sets that were freely available for sharing. Nor had any symbols been evaluated for transparency or cultural sensitivity by local AAC users, their supporting professionals and families.

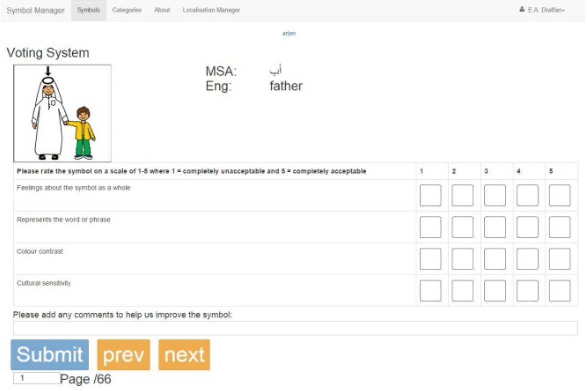
A bespoke Symbol Management system was developed that allowed the team to store symbols. The system also offered participants the chance to take an active role in the decisions made around the development and evaluation of appropriate symbols as they could see and vote on uploaded symbols representing the core vocabularies previously collected.

The online database was based on a Model-View-Controller (MVC) framework using MongodB with JavaScript (NodeJS and an Express JS plugin). The code is open source and available on bitbucket. View templates which generated the html pages were built suing the Jade templating engine. The only other plugins used were for authentication and list filtering. The latter will provide the basis for browse and search features in the final Arabic Symbol Dictionary website.

4.1. Building symbol acceptance system

As part of the online management system a simple voting set up was created using the filters developed for batches of symbols. During voting sessions participants have been presented with a series of around 60-65 images of newly designed symbols, the referent in MSA, Qatari (where applicable) and English. The voting criteria are presented with large selection areas on a scale of 1 to 5 where 5 is completely acceptable (see Figure 3) so that different visual displays can be used. The four criteria are listed with a free text box for comments:

- Feelings about the symbol as a whole
- Represents the word or phrase
- Color contrast
- Cultural sensitivity



Voting System

MSA: أب
Eng: father

Please rate the symbol on a scale of 1-5 where 1 = completely unacceptable and 5 = completely acceptable	1	2	3	4	5
Feelings about the symbol as a whole	□	□	□	□	□
Represents the word or phrase	□	□	□	□	□
Colour contrast	□	□	□	□	□
Cultural sensitivity	□	□	□	□	□

Please add any comments to help us improve the symbol:

1 Page /66

Figure 3 *Voting system with criteria for acceptance on a scale of 1-5 where 5 is completely acceptable*

4.2. Results from voting sessions

The initial batch of symbols had 63 voters logging into the Symbol Manager resulting in 2341 votes for 65 symbols. Overwhelmingly the decisions were very favourable with all mean ratings significantly greater than a rating of 3.5. The average was 4.0. (See Table 1) All voting data was anonymized and comments collated to inform the graphic designer.

Two AAC users were also able to vote on the symbols via an adapted system using their own Sensory Software Grid 2 systems with the symbols added plus a 1-5 or 1-3 ‘thumbs up’ to ‘thumbs down’ scoring depending on their ability. This produced equally good results and comments were captured via recordings. More AAC users are being encouraged to join the forum and as further batches of symbols are developed it is hoped that voting sessions will continue to occur both during face to face meetings and remotely.

Table 1. *One Sample T test for Difference of Mean Ratings from 3.5*

Criteria	**Number of voters**	**Mean rating**	**2 tail P Value for difference from 3.5**
1	63	3.94	<0.0001
2	63	3.90	<0.0001
3	63	4.07	<0.0001
4	63	4.10	<0.0001

4.3. Discussion about the Symbol Management system

The initial development of the Symbol Management system was purely for the team to upload lexical entries and symbols with a set of filter systems based on parts of speech, gender, number and symbol descriptions. However, as the participation by AAC users, their families, therapists and teachers grew it became essential to offer a voting system that quickly produced results because specialists wanted to use the symbols as they were developed. As all the speech therapists and teachers involved had worked for several years with AAC users, but were mainly from countries other than Qatar, it was felt that there should be a method to check acceptability within the community before releasing them for download, not just depending on the team’s opinions. The team had already set up a Google+ method for initially evaluating iconicity and transparency [22].

Those therapists working in the Doha area were very willing to express their opinions about symbol suitability and the links with the corresponding word lists collected. It was noted that there was a general understanding that the lexical entries in Modern Standard Arabic and those entries in Qatari colloquial Arabic may share the same symbol for similar meaning words or multiword phrases but there may need to be additional symbols and / or changes in symbol labels to represent different parts of speech, gender and number and to take into account the bilingual nature of the dictionary to aid those who were not fluent Arabic speakers.

5. Conclusion

The core vocabulary and symbol management systems have provided the research team with quick and easy ways to analyse data as well as provide a platform for user participation. Having a selection of MSA and Qatari core and fringe vocabularies has been essential for ongoing symbol development, but there is still a need to continually update the collection of local vocabularies to ensure that colloquial as well as written language is captured. The present frequency levels of the words collected in Doha (List a) are low in comparison to global lists (Lists b). They are also subjective, based on the AAC forum input rather than a wide base of Arabic AAC users and carers. However, with support it has been shown that where suitable core vocabularies are implemented alongside appropriate symbols AAC users, who have the capacity, can enhance their communication and improve their readiness for reading [24] and already in this project AAC users have greeted the newly developed symbols with much appreciation, but there remains the need to 'focus on long-term outcomes' [9].

There remains the debate as to the differences in parts of speech seen in English core vocabulary lists compared to some Arabic lists with high levels of noun use. It is important to appreciate the limitations of the collection procedures as well as the problems of automated comparisons between lists that require normalization and have different methods for showing root words, different parts of speech and verb declensions.

There is much research still to be carried out to ensure that an appropriate vocabulary list suitable for Arabic AAC users and the development of literacy skills can be collated in a diglossia situation. But as an increasing number of words lists are provided by participants set against the further analysis of the frequency lists already gathered it is felt that this can be achieved.

6. Acknowledgements

This research was made possible by the NPRP award [NPRP 6 - 1046 - 2 - 427] from the Qatar National Research Fund (a member of The Qatar Foundation) and thanks must go to all those participants in Doha who have contributed to the work of the Arabic Symbol Dictionary team. Grateful thanks are also expressed to the ARASAAC team for allowing their symbols to be used with participants. The statements made herein are solely the responsibility of the authors.

7. References

[1] M. B. Huer, "Culturally inclusive assessments for children using augmentative and alternative communication (AAC)," *Journal of Children's Communication Development,* 19 (1), 23–34. 1997.

[2] M. B. Huer, "Examining perceptions of graphic symbols across cultures: Preliminary study of the impact of culture/ethnicity," *Augmentative and Alternative Communication* 16 (3): 180–185. 2000. doi:[10.1080/07434610012331279034]

[3] S. Balandin and T. Iacono, "A few well-chosen words," *Augmentative and Alternative Communication,* 14(September), 147–161 1998.

[4] M. Banajee, C. Dicarlo, and S. Buras Stricklin, "Core Vocabulary Determination for Toddlers," *Augmentative and Alternative Communication,* 19(2), 67–73. 2003.

[5] M. Lahey, and L. Bloom, "Planning a First Lexicon: Which Words to Teach First," *Journal of Speech and Hearing Disorders*, 340–351 1975.

[6] G. M. Van Tatenhove, "Building Language Competence With Students Using AAC Devices: Six Challenges," *Perspectives on Augmentative and Alternative Communication,* 18(2), 38–47 2009.

[7] G. C. Vanderheiden, and D. P. Kelso, "Comparative analysis of fixed-vocabulary communication acceleration techniques," *AAC Augmentative and Alternative Communication,* 3, 196-206. 1987.

[8] M. Lundälv and S. Derbring, "AAC Vocabulary Standardisation and Harmonisation," *Springer-Verlag Berlin Heidelberg,* pp.303–310. 2012.

[9] J. Light, and D. Mcnaughton, "Designing AAC Research and Intervention to Improve Outcomes for Individuals with Complex Communication Needs," *Augmentative and Alternative Communication, (ahead-of-print),* 1-12. 2015.

[10] R. Patel and R. Dakwar-Khamis, "An AAC training program for special education teachers: A case study of Palestinian Arab teachers in Israel," *Journal of Augmentative and Alternative Communication,* 21, 3, 205-217. 2005.

[11] S. Abu-Rabia, "Learning to read in Arabic: Reading, syntactic, orthographic and working memory skills in normally achieving and poor Arabic readers," *Reading Psychology: An International Quarterly,* 16, 351–394. 1995.

[12] S. Abu Rabia, D. Share and S. M. Mansour, "Word recognition and basic cognitive processes among reading-disabled and normal readers of Arabic," *Reading and Writing: An Interdisciplinary Journal*, 16, 423-442. 2003. doi:[10.1023/A:1024237415143]

[13] S. Uziel-Karl, F. Kanaan, R. Yifat, I. Meir, N. Abugov, and D. Ravid, "Hebrew and Palestinian Arabic in Israel: Linguistic Frameworks and Speech-Language Pathology Services," *Topics in Language Disorders* Vol 34 Number 2, p 133 – 154 2014.

[14] B. Woll, and S. Barnett, "Toward a Sociolinguistic Perspective on Augmentative and Alternative Communication," *AAC Augmentative and Alternative Communication,* 14(December), pp.200–211. 1998.

[15] K.J. Hill, and C. Dollaghan, "Conversations of Three-Year Olds: Implications for AAC Outcomes," *American Speech-Language-Hearing (ASHA) Convention. San Francisco, CA.* November. 1999.

[16] W. Zaghouani, "Critical Survey of the Freely Available Arabic Corpora," *In the Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Rejkavik, Iceland,* 26-31 May 2014.

[17] A. Kilgarriff, F. Charalabopoulou, M. Gavrilidou, J. B. Johannessen, S. Khalil, S. J. Kokkinakis and Volodina, E. "Corpus-based vocabulary lists for language learners for nine languages," *Language Resources and Evaluation,* 1-43 2013.

[18] W. Zaghouani, B. Mohit, N. Habash, O.Obeid, N. Tomeh, and K. Oflazer. "Large-scale Arabic Error Annotation: Guidelines and Framework," *In the Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014). Rejkavik, Iceland,* 26-31 May 2014.

[19] A. Oweini and K. Hazoury, "Towards a list of Awards a Sight Word List in Arabic," *International Review of Education,* 56 (4), 457-478 2010.

[20] K. Hill, and B. Romich, *100 Frequently Used Core Words.* Accessed May 2015 https://aaclanguagelab.com/files/100highfrequencycorewords2.pdf

[21] K. Hill, and B. Romich, "A summary measure clinical report for characterizing AAC performance," *Proceedings of the RESNA '01 Annual Conference, Reno, NV.* pp 55-57. 2001.

[22] J. Boenisch and G. Soto, "The Oral Core Vocabulary of Typically Developing English-Speaking School-Aged Children," *Implications for AAC Practice. Augmentative and Alternative Communication,* pp.77–84. 2015.

[23] T. Buckwalter and D. Parkinson, "A frequency dictionary of Arabic: Core vocabulary for learners," Routledge. 2014.

[24] D. Evans, L. Bowick, M. Johnson and P. Blenkhorn, "Using iconicity to evaluate symbol use," *In:*

Proceedings of the 10th international conference on computers helping people. Linz, Austria, pp 874–881 2006.

[25] P. Hatch, L. Geist, and K. Erickson, “Teaching Core Vocabulary Words and Symbols to Students with Complex Communication Needs,” *Presented at Assistive Technology Industry Association,* 2015. Retrieved 19/2/2015 fromhttp://www.med.unc.edu/ahs/clds/files/conference-hand-outs/atia_2015.pdf (Accessed 14 June 2015).

Remote Speech Technology for Speech Professionals - the CloudCAST Initiative

Phil Green[1], *Ricard Marxer*[1], *Stuart Cunningham*[1], *Heidi Christensen*[1],
Frank Rudzicz[2,3], *Maria Yancheva*[3], *André Coy*[4],
Massimiliano Malavasi[5], *and Lorenzo Desideri*[5]
[1]University of Sheffield, United Kingdom,
[2]Toronto Rehabilitation Institute, Canada, [3]University of Toronto, Canada,
[4]University of West Indies, Jamaica, [5]AIAS Onlus Bologna, Italy

Abstract

Clinical applications of speech technology face two challenges. The first is data sparsity. There is little data available to underpin techniques which are based on machine learning and, because it is difficult to collect disordered speech corpora, the only way to address this problem is by pooling what is produced from systems which are already in use. The second is personalisation. This field demands individual solutions, technology which adapts to its user rather than demanding that the user adapt to it. Here we introduce a project, CloudCAST, which addresses these two problems by making remote, adaptive technology available to professionals who work with speech: therapists, educators and clinicians.

Index Terms: assistive technology, clinical applications of speech technology

1. Introduction to CloudCAST

In this working paper, we introduce CloudCAST, a Leverhulme Trust International Network funded from January 2015 for 3 years. The network partners are The University of Sheffield (United Kingdom), AIAS Onlus Bologna (Italy), The University of the West Indies (Jamaica), and the University of Toronto (Canada).

In recent years, there has been significant progress in Clinical Applications of Speech Technology (CAST) in diagnosis of speech disorders [1], tools to correct pronunciation and improve reading skills [2], recognition of disordered speech [3] and voice reconstruction by synthesis [4]. The aim of CloudCAST is to make progress in this domain and to provide a freely-available platform for worldwide collaboration.

We aim to place CAST tools in the hands of professionals who deal with clients with speech and language difficulties, including therapists, pathologists, teachers, and assistive technology experts. We intend to do this by means of a free-of-charge (if possible), remotely-located, internet-based resource ‘in the cloud’ which will provide a set of software tools including personalised speech recognition, diagnosis and interactive spoken language learning. Following a user-centred design methodology, we will provide interfaces which will make these tools easy to use for professionals and their clients, who are not necessarily speech technology experts.

There are various models for user-centred design [5], among which the ISO standard 9241-210 [6] is prominent. This standard for human-centred design processes includes six guiding principles (P): P1. understand the user, the task and environmental requirements; P2. encourage early and active involvement of users; P3. be driven and refined by user-centered evaluation; P4. include iteration of design solutions; P5. address the whole user experience; P6. encourage multi-disciplinary design.

The CloudCAST resources will also facilitate speech data collection necessary to inform the machine learning techniques which underpin this technology: we will be able to automatically collect data from systems which are already in use, as well as provide a database scheme for collecting and hosting databases related to this domain. Our 3-year aim is to create a self-sustaining CloudCAST community to manage future development beyond our current funding period.

While CloudCAST will build on previous work by its partners and others, we believe that it offers several ‘unique selling points’, including:

- The resource will be available worldwide, and free of charge.
- We will provide interfaces, resources and tools targeted at several kinds of users, including:
  - Developers, who want to embed CloudCAST technology into their own applications, for instance voice control of domestic robots,
  - Speech professionals, who want to use CloudCAST technology to work with their clients, for instance, to devise personalised therapy exercise programmes,
  - End users, for whom applications are developed, e.g., children learning to read,
  - Speech technologists, who are improving or adding to the CloudCAST technology itself.
- The technology will be based on open source toolkits such as Kaldi for automatic speech recognition and OpenHab for smart homes [7, 8].
- Subject to ethical constraints, we will collect speech data and metadata from every CloudCAST interaction. All this material will therefore be available for re-training the technology, and for analysis. In this way,
  - we will be able to personalise the technology for each End User,
  - by pooling the data, we will address the problem that for abnormal speech the large datasets needed for speech technology development are not available,
  - we will be able to underpin and evaluate improvements in analysis and classification of speech disorders.

2. Challenges for CloudCAST

CloudCAST's success requires meeting a number of technical, scientific and more general challenges:

- The technology will run remotely, but in many applications it must deliver results rapidly, within a few seconds.
- The technology should improve its performance as it is used, by adaptation to the data it is collecting.
- It will not be possible to control the conditions under which the tools are used to the extent that one might like. For example, diverse recording devices and recording conditions may make normalization challenging.
- There must be shared functionality of tools over applications. For instance, pronunciation tutors and reading tutors have much in common.
- There must be interfaces, and guides to these interfaces, which are suitable for each user-group listed above.
- There must be a scheme which protects the security and privacy of CloudCAST users and their data.
- There is understandable resistance to technology from some speech professionals, based on bad experiences.
- For this reason, and others, the technology must adapt to its user, rather than the other way round.
- There must be a strategy for developing a self-sustaining CloudCAST community.

Our intention is to commence with three exemplar applications: small vocabulary command-and-control with disordered speech, a literacy tutor and a computer aid for therapists. These are described after the next section, in which we introduce the common speech technology resource that will support them.

3. Speech technology resource

Several toolkits exist which provide core speech recognition facilities on which applications can be built, notably Speechmatics [9], Google's Web Speech API [10] and SoundHound [11]. Speechmatics provides a queue-based speech transcription service supporting multiple languages and audio formats, performs automatic punctuation, capitalisation and diarization (speaker separation) and supplies individual word timings and confidences. It's authors claim to achieve near real-time turnaround with very high accuracy. Google, through its proposed Web Speech API, provides both speech recognition and synthesis. The speech recognition service outputs the results in the form of multiple hypotheses of word-level transcriptions with associated confidence scores. SoundHound provides a speech-to-meaning service that performs simultaneous speech recognition and natural language understanding. This process outputs its results in the form of structured commands instead of plain text transcription.

For CloudCAST, these solutions fall short in terms of the types and details of the results they return, the flexibility of the recognition process, provisions for customisation of the speech models, and modes of interaction. The maximum level of detail provided in all these solutions is an N-best list of word-level transcriptions with associated confidences. In the case of Speechmatics, word-level time alignments are also available. However higher-level details such as phone time alignments are not accessible. Furthermore, other types of results such as decoding lattices and word confusion networks (WCN) [12] are not provided. The grammars (or language models) used in these systems are fixed to general-domain dictation applications (in Google's Web Speech API, the introduction of a grammar specification function was discussed in 2012, but to the best of our knowledge it has not been concretized or implemented in Chrome). While Googles service does provide an interactive mode in which partial results of the decoding process are immediately available, this is not the case with the service provided by Speechmatics. None of the services provide any means of creating custom models using specific training material. This precludes targeting disordered speech or other niche cases.

The requirements of CloudCAST include providing an interactive speech recognition service where the client must be able to modify the grammar, the model, and other relevant parameters. The client should have instant feedback about the recognition process, such as partial decoding as well as access to fully detailed results such as phone-level alignments and posterior probabilities. Crucially, interactions of clients with CloudCAST should provide data resources to improve the recognition process and the training of future models.

The main architecture of CloudCAST (Figure 1) can be split into the exemplars, the frontend, and the backend. The exemplars are services using CloudCAST, for instance, webapps that perform literacy tutoring or command-and-control (see next section). The frontend is the visible CloudCAST website, from which users can manage their recordings, developers can obtain API keys, professionals can create models, and so on. Finally, the backend is the server which consumes audio from the exemplars and provides speech recognition results. The backend is also in charge of applying the parameter changes that the exemplars may request to the recognition process.

Both the frontend and the backend have access to a common storage space and database where they store models, recordings, and authentication details. The frontend and backend are both backed by worker processes, whose roles are to perform computationally intensive tasks, such as the training of the models and actual speech recognition, which may be run in separate devices. This split ensures the scalability of the system.

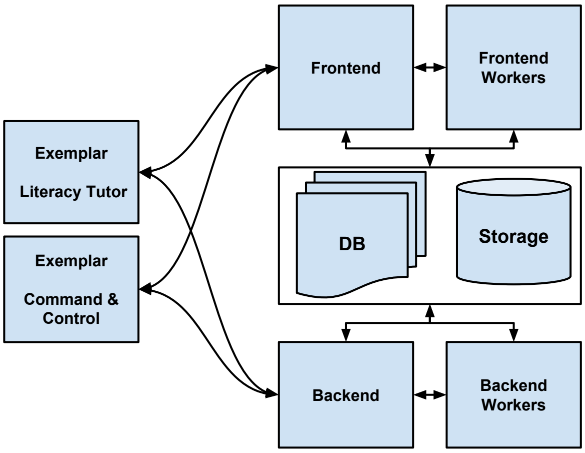


Figure 1: Architecture of the CloudCAST resource.

CloudCAST is developed with open source software. We have decided to create the frontend using Flask, a free software microframework for web development.

To implement the speech recognition service (the backend) we have decided to build on kaldi-gstreamer-server developed by Tanel Alume [13]. Kaldi-gstreamer-server is a distributed online speech-to-text system that features real-time speech recognition and a full-duplex user experience where the partially transcribed utterance is provided to the client. The system includes a simple client-server communication protocol and scalability to many concurrent sessions. The system is open source and based on free software and therefore serves as a starting point for building CloudCAST, allowing us to deploy recognisers developed at Sheffield within the CloudCAST framework [14]. It uses Kaldi [7] for speech recognition processing. Kaldi is a well-known free software library widely used in the research community partly due to its modular and flexible architecture.

To facilitate the creation of services using CloudCAST, we are also developing a speech recognition client in JavaScript based on the existing library dictate.js. The proposed client extends dictate.js with multiple types of interactions with the server, such as swapping grammars, models and other parameters, as well as interpreting the different results provided by the server.

4. Exemplars

4.1. Literacy tutor

Among the first set of exemplars to be developed will be an automated literacy tutor. In some respects the literacy tutor represents the most complex type of application that can be developed using the tools CloudCAST will make available. In addition to being a good showcase for the tools, the literacy tutor can be useful in bolstering current efforts to combat illiteracy as it will be a freely available, cloud-based resource that can be modified to meet the needs of individual users.

It has been shown that the use of speech-enabled literacy tutors can lead to significant improvement in their users ability to read [15, 16, 17]. Among the best known systems is Project LISTEN [18]. Project LISTEN, developed at Carnegie Mellon University, works as a tutor by listening to, as well as reading to the user. This system was one of the first to employ feedback that was able to effectively respond to readers when they encountered challenges or made mistakes. When the system was deployed in schools it was found that students using the reading tutor outperformed their peers who learned from regular classroom-based activities and compared favourably to students given one-to-one tutoring by human experts. Outside of the United States, the effectiveness of Project LISTEN to provide tutorial support for language learners has been tested (on very limited scales) in a number of countries, including Canada, Ghana and India [19, 20, 21]. Users were shown to make significant progress in literacy skills when they used the tutor.

Project LISTEN is available for research purposes, but it is not a commercial product; it is not openly accessible, nor is it cloud-based. One commercial, web-based reading tutor is the Reading Trainer component of the Ridinet online network. Ridinet is meant to provide practice and training in literacy and numeracy for Italian children diagnosed with autism spectrum disorder [22]. Reading Trainer was included in the system for the purpose of increasing reading fluency. The initial phase of the Reading Trainer is a speed test, where users are prompted to read a previously unseen text and the time it takes them to read it is used as their initial reading speed. The tool is customisable; it allows the user to select, among other parameters, reading speed, reading accuracy, reading unit and story length. The level of feedback can also be set to either prompt the user or praise their performance and effort.

The basic functionality of the CloudCAST literacy tutor exemplar will be similar to the two tutors described above, but will differ in at least three important respects Firstly, it will be freely available to anyone with an internet connection. Secondly, the tool will be further customisable: the user will be able to change language and upload new stories. Finally, the tutor will have an integrated reading age assessment tool to determine the reading level of the user and to act as a pre-test for potential learning challenges. The results of assessment and user performance for each session will be securely stored online for easy tracking of their progress.

4.2. Environmental control

The command and control exemplar will provide a service that will allow, for instance, manipulation of multiple devices in a smart home either directly with speech commands or through voice communication with assistive robots.

Current home automation systems and the increasingly popular Internet of Things (IoT) can provide great support to people with disabilities by improving their autonomy and safety in daily living activities.

There are several ways in which CloudCAST will improve on existing speech recognition solutions. Although user interfaces based on voice interaction are particularly suited for this type of application, current systems devised for assistive technology or for the mainstream market are unsuitable, in terms of performance, for many potential users. Common limitations are the inability to be completely hands-free and poor recognition performance.

Command and control systems are particularly useful for subjects with mobility issues. In many cases these people also experience speech disorders for which available speech recognition systems are not optimized. The possibility of using personalised speech models could greatly enhance the recognition accuracy and therefore the reliability of the system. Furthermore the speech material produced by such users will be of great value to improve future speech models for other users with similar issues.

The ability to define a customised grammar will render the system significantly more robust to speech disfluencies, environment noise and recognition ambiguity. Keeping grammars simple, with few word options (low perplexity) at each stage of the control sequence will make the system less prone to recognition errors.

Since many actual home automation fieldbuses can be easily connected to the internet and IoT devices are natively equipped with this property, cloud services developed within the CloudCAST project can be easily implemented, and will be flexible and customizable. The potential of the exemplars can be extended through the use of specific open source servers dedicated to the integration of home automation technologies and IoT solutions, such as Openhab [8].

4.3. Speech therapy

The ability to communicate is one of the most basic human needs. Many lose the ability to communicate, due to a range of health conditions which result in a speech impairment. Speech therapy helps improve communication ability and produces benefits in terms of quality of life and participation in society. Articulation therapy aims to improve the speech of people with speech impairment. It is however time-consuming, and patients rarely receive sufficient therapy to maximise their communication potential [23, 24].

In articulation therapy speech therapists work with patients on the production of specific speech sounds and provide feedback on the quality of these speech sounds. This process helps the patient improve their production of these sounds thereby improving the overall intelligibility of speech. Our previous research shows that computer programs using speech recognition can improve outcomes of speech therapy for adults with speech difficulties [25, 26]

For our CloudCAST exemplar we intend to build on our past work to develop a web-based application. This demonstrator will enable therapists and clients to work together to specify speech exercises. These exercises could then be independently completed by the client between therapy sessions.

A big advantage of using technology over traditional practice will be that therapists can monitor and review the progress that their client has made. During the completion of the exercises, the speech produced by the client will be scored and then stored for review. This means that any difficulties that they encounter during the exercises can be identified and discussed with the therapist.

We have previously developed techniques for using automatic speech recognition to provide feedback to patients practising their speech [25, 27]. These approaches are based on using specially developed speech recognition software able to provide objective feedback, which acts as a substitute for the judgement of an expert listener, such as the speech and language therapist. This feedback can be given to patients when they are practising either with a therapist or on their own between therapy sessions [26]. We will use especially adapted recognisers available via the CloudCAST platform to generate this objective feedback. We will then make these approaches available in a range of motivational exercises.

5. Data collection and repository

CloudCAST will also serve as a data repository for the distribution of existing databases and for the acquisition of new databases, along with provided tools for that collection. Below we discuss the first database that will become freely available in CloudCAST, TORGO, and the database scheme we will use to represent future data collection

5.1. TORGO

TORGO consists of aligned acoustic and EMA measurements from individuals with and without cerebral palsy (CP), each of whom recorded 3 hours of data [28]. CP is one of the most prevalent causes of speech disorder, and is caused by disruptions in the neuro-motor interface [29] that do not affect comprehension of language, but distort motor commands to the speech articulators, resulting in relative unintelligibility [30]. The motor functions of each participant in TORGO were assessed according to the standardized Frenchay Dysarthria Assessment [31] by a speech-language pathologist affiliated with the Holland-Bloorview Kids Rehab hospital and the University of Toronto. Individual prompts were derived from non-words (e.g., /iy-p-ah/ [32]), short words (e.g., contrasting pairs from [33]), and restricted sentences (e.g., the sentence intelligibility section in the Yorkston-Beukelman Assessment [34], and sentences from MOCHA-TIMIT.

The EMA data in TORGO were collected using the three-dimensional Carstens Medizinelektronik AG500 system [35, 36]. Sensors were attached to three points on the surface of the tongue, namely tongue tip (TT – 1 cm behind the anatomical tongue tip), the tongue middle (TM – 3 cm behind the tongue tip coil), and tongue back (TB, approximately 2 cm behind the tongue middle coil). A sensor for tracking jaw movements (JA) was attached to a custom mould over the lower incisors [37]. Four additional coils were placed on the upper and lower lips (UL and LL) and the left and right corners of the mouth (LM and RM). Reference coils were placed on the subject's forehead, nose bridge, and behind each ear above the mastoid bone.

5.2. Future database scheme

New users of CloudCAST can immediately use our database framework for representing the data. To a large extent, this framework is designed to be generic to all speech recording tasks, and not all components need to be utilized. The database schema is broken down into three core sections: the subject, the task, and the session. A high-level overview of the data representation is shown in Figure 2.

The subject section generally involves aspects related to the speaker, including demographics, levels of permission to use the data, and factors affecting the subject's language quality, such as country of origin, country of residence, spoken languages, history of smoking, and education level. The task section specifies the language task (e.g., picture description, conversation, reading of text, repetition of audio) along with a bank of available task instances (e.g., pictures to be used in the picture description task). The system supports a variety of question and answer types, including text, speech, multiple-choice, and fill-in-the-blank, with the ability for easy extension to new types. Each task instance is optionally rated with a level of difficulty, measured across arbitrary dimensions (e.g., phonological complexity, syntactic complexity). Information related to automatic scoring of tasks is stored along with each task instance, where appropriate (e.g., the correct answer to a multiple-choice question). Each subject can be associated with a number of recording sessions, and each session can be associated with a number of task instances. The session section stores the subject responses to specific task instances every time they interact with the system. This includes their language data, as well as metadata such as total amount of time spent on each task, and date of completion.

This database is designed to be extensible to future needs, and will be especially useful to streamline data organization to projects that otherwise have a more clinical focus. It enables (i) longitudinal subject assessments, due to the ability to accommodate multiple language task instances in order to avoid 'the learning effect' over time, (ii) dynamic variation of task instance difficulty and type based on subject performance, and (iii) automated scoring of subject performance where appropriate.

5.3. Ethics

As part of the CloudCAST initiative we will be seeking to collect speech data from individual participants. To do so we must ensure that we fully respect their personal data. As part of this process professionals who initiate a service through CloudCAST will need to first confirm that they are abiding by the local ethics and governance rules.

For individual participants making use of CloudCAST services we will follow a process approved by the University of Sheffield Research Ethics Committee. It is proposed that as part of this process we will first fully explain to each individual user when they register with CloudCAST the background to the

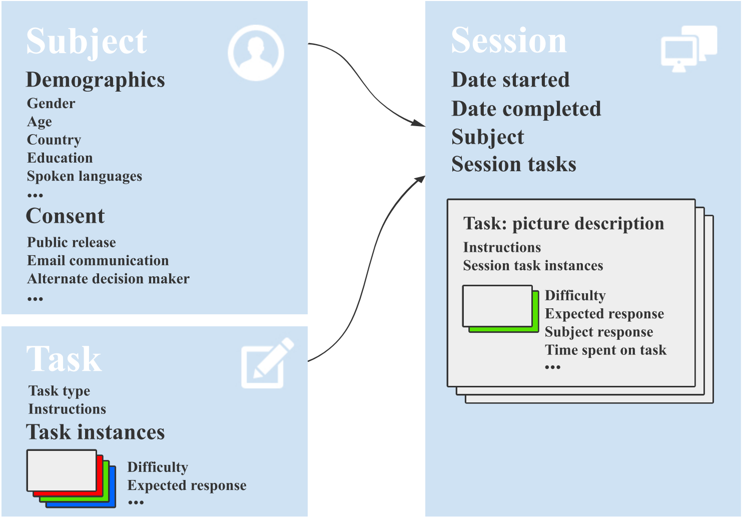


Figure 2: *Simplified database schema, arranged into three core sections: Subject, Session, and Task.*

project and how we intend to use their speech data. Participants will be able to opt-in to different levels of engagement with the CloudCAST initiative. At the most basic level, a participant will be able to make use of the CloudCAST services without their data being used for further research, or shared with other researchers. The second level of participation can be selected by the participant when they wish to allow the CloudCAST team to retain their data for further research. The final level of participation can be selected by participants when they wish their data to be retained and potentially distributed to other speech researchers.

As part of the on-going relationship with the participants, they will periodically be asked to re-confirm their consent for their data to be used in the way they chose.

6. Conclusions

CloudCAST aims to create a self-sustaining community of academic and speech professionals which will continue to grow after its 3 year funding period. It is our belief that only by collaborating in this way can we make the benefits of speech technology available to those who need it most and at the same time create the knowledge bases for further technical improvement. To attain critical mass we need to widen the participants beyond the initial partners. If you are interested, please contact us by registering on our website: http://cloudcast.rcweb.dcs.shef.ac.uk/

7. Acknowledgements

CloudCAST is an International Network (IN-2014-003) funded by the Leverhulme Trust.

8. References

[1] “PEAKS a system for the automatic evaluation of voice and speech disorders,” *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[2] O. Saz, S.-C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodrguez, “Tools and technologies for computer-aided speech and language therapy,” *Speech Communication*, vol. 51, no. 10, pp. 948–967, 2009.

[3] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, “A comparative study of adaptive, automatic recognition of disordered speech,” in *Proc Interspeech 2012*, Portland, Oregon, US, Sep 2012.

[4] C. Veaux, J. Yamagishi, and S. King, “Towards personalized synthesized voices for individuals with vocal disabilities: Voice banking and reconstruction,” in *Proceedings of 4th Workshop on Speech and Language Processing for Assistive Technologies, SLPAT2013*, 2013, pp. 107–111.

[5] S. Blackburn and P. Cudd, “An overview of user requirements specification in ICT product design,” in *Proceedings of the AAATE workshop: The social model for AT Technology Transfer*, Sheffield, UK, 2010.

[6] “Iso 9241-210: 2009. ergonomics of human system interaction - part 210: Human-centrered design for interactive systems,” Switzerland, 2009.

[7] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[8] “openHAB.” [Online]. Available: http://www.openhab.org

[9] “Speechmatics.” [Online]. Available: https://www.speechmatics.com/

[10] G. Shires and H. Wennborg, “Web speech api specification,” W3C, Community Group Final Report, Oct. 2012, https://dvcs.w3.org/hg/speech-api/raw-file/9a0075d25326/speechapi.html.

[11] “Soundhound - houndify platform.” [Online]. Available: http://www.soundhound.com/houndify

[12] D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tur, “Beyond asr 1-best: Using word confusion networks in spoken language understanding,” *Computer Speech & Language*, vol. 20, no. 4, pp. 495–514, 2006.

[13] T. Alumäe, “Full-duplex speech-to-text system for Estonian,” Kaunas, Lihtuania, 2014.

[14] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, “Automatic selection of speakers for improved acoustic modelling : Recognition of disordered speech with sparse data,” in *Spoken Language Technology Workshop, SLT’14*, Lake Tahoe, Dec 2014.

[15] J. Mostow, G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, M. B. Sklar, and B. Tobin, “Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction,” *Journal of Educational Computing Research*, vol. 29, pp. 61–117, 2003.

[16] M. J. Adams, “The promise of automatic speech recognition for fostering literacy growth in children and adults,” in *Handbook of Literacy and Technology*, M. McKenna, L. Labbo, R. Kieffer, and D. Reinking, Eds. Hillside, New Jersey: Lawrence Erlbaum Associates, 2005, vol. 2, pp. 109–128.

[17] B. Wise, R. Cole, S. van Vuuren, S. Schwartz, L. Snyder, N. Ngampatipatong, J. Tuantranont, and B. Pellom, “The promise of automatic speech recognition for fostering literacy growth in children and adults,” in *Interactive literacy education: Facilitating literacy environments through technology*, C. Kinzer and L. Verhoeven, Eds. Mahwah, New Jersey: Lawrence Erlbaum, 2005, vol. 2, pp. 31–76.

[18] J. Mostow, S. Roth, A. Hauptmann, and M. Kane, “A prototype reading coach that listens,” in *Association for the Advancement of Artificial Intelligence, AAAI-94*, Seattle, Washington, 1994.

[19] T. Cunningham, “The effect of reading remediation software on the language and literacy skill development of ESL students,” Master’s thesis, University of Toronto, Toronto, Canada, 2006.

[20] G. Korsah, J. Mostow, M. Dias, T. Sweet, S. Belousov, M. Dias, and H. Gong, “Improving child literacy in africa: Experiments with an automated reading tutor,” *Information Technologies and International Development*, vol. 6, pp. 1–19, 2010.

[21] F. Weber and K. Bali, “Enhancing esl education in india with a reading tutor that listens,” in *Proceedings of First ACM Symposium on Computing for Development ACM*, vol. 20, London, UK, 2010, pp. 1–9.

[22] S. Pinnelli, “Dyslexia and young adults. A case study: from assessment to intervention with reading trainer software,” in *Proceedings of SIREM 2013*, Bari, Italy, 2014, pp. 84–94.

[23] J. Law, Z. Garrett, and C. Nye, “Speech and language therapy interventions for children with primary speech and language delay or disorder,” *Cochrane Database of Systematic Reviews*, no. 3, 2003.

[24] P. Enderby and L. Emerson, *Does Speech and Language Therapy Work?* London: Singular, 1995.

[25] R. Palmer, P. Enderby, and S. P. Cunningham, “Effect of three practice conditions on the consistency of chronic dysarthric speech,” *Journal of Medical Speech-Language Pathology*, vol. 12, no. 4, pp. 183–188, 2004.

[26] R. Palmer, P. Enderby, and M. Hawley, “Addressing the needs of speakers with longstanding dysarthria: computerized and traditional therapy compared.” *International journal of language & communication disorders / Royal College of Speech & Language Therapists*, vol. 42 Suppl 1, pp. 61–79, Mar. 2007.

[27] M. Parker, S. P. Cunningham, P. Enderby, M. S. Hawley, and P. D. Green, “Automatic speech recognition and training for severely dysarthric users of assistive technology: the STARDUST project,” *Clinical Linguistics & Phonetics*, vol. 20, no. 2-3, pp. 149–156, 2006.

[28] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,” *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[29] R. D. Kent and K. Rosen, “Motor control perspectives on motor speech disorders,” in *Speech Motor Control in Normal and Disordered Speech*, B. Maassen, R. Kent, H. Peters, P. V. Lieshout, and W. Hulstijn, Eds. Oxford: Oxford University Press, 2004, ch. 12, pp. 285–311.

[30] R. D. Kent, “Research on speech motor control and its disorders: a review and prospective,” *Journal of Communication Disorders*, vol. 33, no. 5, pp. 391–428, 2000.

[31] P. M. Enderby, *Frenchay Dysarthria Assessment*. College Hill Press, 1983.

[32] J. W. Bennett, P. Van Lieshout, and C. M. Steele, “Tongue control for speech and swallowing in healthy younger and older subjects,” *International Journal of Orofacial Myology*, vol. 33, pp. 5–18, 2007.

[33] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, “Toward phonetic intelligibility testing in dysarthria,” *Journal of Speech and Hearing Disorders*, vol. 54, pp. 482–499, 1989.

[34] K. M. Yorkston and D. R. Beukelman, *Assessment of Intelligibility of Dysarthric Speech*. Tigard, Oregon: C.C. Publications Inc., 1981.

[35] A. Zierdt, P. Hoole, and H. G. Tillmann, “Development of a system for three-dimensional fleshpoint measurement of speech movements,” in *Proceedings of the 14th International Conference of Phonetic Sciences (ICPhS’99)*, San Francisco, USA, August 1999.

[36] P. Van Lieshout, G. Merrick, and L. Goldstein, “An articulatory phonology perspective on rhotic articulation problems: A descriptive case study,” *Asia Pacific Journal of Speech, Language, and Hearing*, vol. 11, no. 4, pp. 283–303, 2008.

[37] P. H. Van Lieshout and W. Moussa, “The assessment of speech motor behavior using electromagnetic articulography,” *The Phonetician*, vol. 81, pp. 9—22, 2000.

Speech and language technologies for the automatic monitoring and training of cognitive functions

Anna Pompili, Cristiana Amorim, Alberto Abad, Isabel Trancoso

L^2F - Spoken Language Systems Lab, INESC-ID Lisboa
IST - Instituto Superior Técnico, University of Lisbon
{anna,cristiana.amorim,alberto,imt}@l2f.inesc-id.pt

Abstract

The diagnosis and monitoring of Alzheimer's Disease (AD), which is the most common form of dementia, has been the motivation for the development of several screening tests such as Mini-Mental State Examination (MMSE), AD Assessment Scale (ADAS-Cog), and others. This work aims to develop an automatic web-based tool that may help patients and therapists to perform screening tests. The tool was implemented by adapting an existing platform for aphasia treatment, known as Virtual Therapist for Aphasia Treatment (VITHEA). The tool includes the type of speech-related exercises one can find in the most common screening tests, totalling over 180 stimuli, as well as the Animal Naming test. Its great flexibility allows for the creation of different exercises of the same type (repetition, calculation, naming, orientation, evocation, ...). The tool was evaluated with both healthy subjects and others diagnosed with cognitive impairment, using a representative subset of exercises, with satisfactory results.

1. Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease which represents 60 to 70% of the dementia cases in Portugal [1, 2, 3]. However, its first signs can go unnoticed [1, 2, 4, 5]. Typically, AD is known to cause alterations of memory and of spacial and temporal orientation [3, 4, 6]. Furthermore, AD increases dramatically with age and it has no cure. Nevertheless, an early diagnosis may slow down its progression by enabling a more effective treatment [5]. For this purpose, several neuropsychological tests exist in the literature, each targeting different cognitive domains and capabilities. The Mini-Mental State Examination (MMSE) [7] and the AD Assessment Scale - Cognitive subscale (ADAS-Cog) [7] are two of the most popular tests used in Portugal for screening cognitive performance and tracking alterations of cognition over time. They involve the assessment of different capabilities, such as orientation to time and place, attention and calculus, language (naming, repetition, and comprehension), or immediate and delayed recall. Another type of test also commonly applied by therapists in the diagnosis of AD is the Verbal Fluency test [7]. In this test, the patient should produce as many words as he can beginning with a particular letter (phonemic fluency test) or belonging to a particular category, e.g. fruits (semantic fluency test), during 60s. This test is used both in assessing the verbal initiative ability and executive function such as the inhibition ability, the difficulty in switching among tasks, and the perseverance attitude [7]. Typically, the most commonly used versions for the Portuguese population consider the letter "P" for the phonemic version and the "Animal" category for the semantic category. Other tests not so frequently adopted for this population are the Wechsler Adult Intelligent Scale - III (WAIS-III) [7], which provides a measure of general intellectual function in older adolescent and adults, and the Stroop test [7], which is a measure of cognitive control, evaluating how easily a person can maintain a goal in mind while suppressing habitual responses.

Most of these tests include a verbal component provided in response to a visual or spoken stimulus solicited by a therapist. Thus, due to their nature, and the need to continuously monitor the cognitive decline over time, these tests lend themselves naturally to be automated through speech and languages technologies (SLT). A tool including the digitized version of these tests with the possibility of an immediate evaluation through automatic speech recognition could be of valuable support in health care centres. The therapist will have access to an organized archive of tests which could be administered in the traditional way, or remotely, when the physical dislocation of the subject is hampered by logistic constraints or physical disabilities. Recordings and evaluations will be stored and made available for later consultation. On the other hand, research has shown that cognitive skills, which can fade without stimulation as we age, can be improved by playing games that stimulate brain activity [8]. An automated tool for the monitoring of AD could be easily extended to support exercises and brain games for the daily training of cognitive capabilities such as short-time memory, attention, calculus, reasoning ability and many others.

Up to our knowledge, there are few works in the literature that exploit SLT to automate certain types of neuropsychological tests. Some of the most relevant are the kiosk system designed to use at home as a prevention instrument for early detection of AD described in [5], the end-to-end system for automatically scoring the Logic Memory test of the WAIS-III presented in [9], and the system that implements a modified version of the MMSE based on the IBM ViaVoice recognition engine of [10]. These works show the recent increasing interest on this area, but also the long road ahead to support the large variety of existing neuropsychological tests (e.g., some of them are not fully automated).

This work makes a step towards filling this gap by introducing a set of neuropsychological tests for AD intended for the Portuguese population, which were integrated into an automatic web-based system [11]. The system presented in this work extends an on-line platform named VITHEA [12] used for aphasia treatment that incorporates SLT to provide word naming exercises. For this to be possible, the system resorts to a keyword spotting technique which consists of detecting a certain set of words by using a competing background model with the keywords model [13]. This platform is used daily by patients and speech therapists and has received several awards from both the speech and the health-care communities. The success of this platform and its flexibility, that allows to create different ex-

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 103–109,
Dresden, Germany, 11 September, 2015.

ercises, have motivated its use as a foundation for this work. Our first step was the automation of the exercises in MMSE and ADAS-COG that involve speech. The second step was the implementation of the semantic fluency test, starting with the Animal category, also known as Animal Naming test. As explained in the next sections, the automation of such tests have raised several technological challenges, both for the automatic speech recognition and text-to-speech synthesis technologies.

In the following, Section 2 briefly presents the VITHEA platform that was used as a foundation for this work, while Section 3 describes the extended system resulting from the implementation of the selected neuropsychological tests into the VITHEA platform. Section 4 reports how each type of test was concretely implemented. Then, in Section 5 the focus is on the experiments, both detailing the automatic speech recognition module, the speech corpus used for evaluation and the experimental results. Finally, Section 6 presents the conclusions and future work.

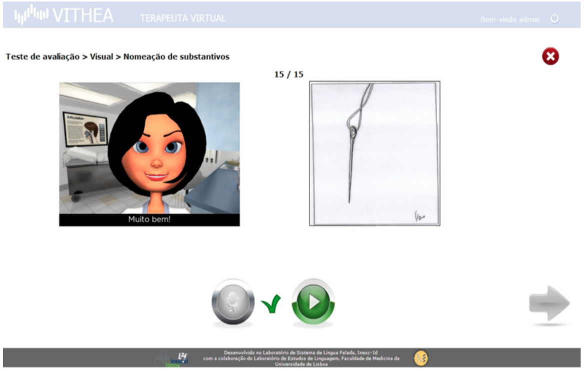
2. The VITHEA platform

VITHEA (Virtual Therapist for Aphasia Treatment) is a web-based platform developed with the collaboration of the Spoken Language Processing Lab of INESC-ID (L^{2}F) and the Language Research Laboratory of the Lisbon Faculty of Medicine (LEL). The system aims at acting as a ”virtual therapist”, allowing the remote rehabilitation from a particular language disorder, aphasia. For this to be possible, the platform comprises two specific modules, dedicated respectively to the patients, for carrying out the therapy sessions, and to the clinicians, for the administration of the functionalities related to them (e.g., manage patient data, manage exercises, and monitor user performance). In this way, speech therapy exercises created by speech therapists through the clinician module, can be later accessed by aphasia patients through the patient module with a web-browser. During the training sessions, the role of the speech therapist is taken by a ”virtual therapist” that presents the exercises and is able to validate the patients answers.

The overall flow of the system can be described as follows: when a therapy session starts, the virtual therapist shows to the patient, one at a time, a series of exercises. These may include either the presentation of images, the reproduction of short videos or audios, and textual information. The patient is then required to respond verbally by naming the contents of the object or action that is represented. The utterance produced is recorded, encoded and sent via network to the server side. Here, a web application server receives the audio file which is processed by the ASR system, generating a textual representation of the patient’s answer. This result is then compared with a set of predetermined textual answers (for the given question) in order to verify the correctness of the patient’s input. Finally, feedback is sent back to the patient with the correctness of the answer provided. Figure 1 illustrates the use of the VITHEA platform.

3. Extending VITHEA for neuropsychological screening

Extending VITHEA for including neuropsychological tests involved important alterations in the original platform, both on the patient and the clinician modules. However, the flexibility of VITHEA allows for the easy addition of new categories of exercises. These can then be combined in multiple ways by the clinician to form new tests, and to create different exercises of the same type. According to the original system, in order to answer the question presented by the virtual therapist, the patient needs to manually interact with the system to start and stop the recording of his answer, and to advance among different stimuli.



Desenvolvido no Laboratório de Sistema de Língua Falada, Inesc-Id
com a colaboração do Laboratório de Estudos de Linguagem, Faculdade de Medicina da
Universidade de Lisboa

Figure 1: A caption of the VITHEA platform during the presentation of an exercise.

The usability of this interface has been adapted to meet the needs of an ageing population, with cognitive impairments. In particular, we considered important to implement the following updates:

- To simplify the interaction with the tool and make the evaluation process more fluid, we minimized the use of the mouse. The interface now automates part of the recording process and the progression between stimuli. An action from the patient is only required to stop the recording process.
- Since cognitive impairments and ageing often results in a limited auditory capability, the speech rate of the therapist has been tuned until finding the best compromise between a more understandable but still natural voice.

Also, the feedback from the neurologists involved in this work provided us important guidelines regarding the presentation of the tests. Following their advices, we introduced the alterations listed below:

- To make the interaction with the system more natural the virtual therapist now provides a random feedback to the patient when the evaluation switches among different classes of stimuli.
- Optional instructions have been added for the more complex questions.
- For some stimuli, the virtual therapist now provides a semantic hint if the patient has not provided an answer after a given amount of time.

The platform now allows also to store some additional personal information of the profile of the patient that are needed for the assessment of some sub-tests (i.e. place of birth, age, etc.), and the result of the assessment in terms of test score obtained. During the application of a neuropsychological test, the scores are individually calculated for each question. After the answer has been processed by the automatic speech recognition (ASR) system, the platform computes both the maximum score allowed for the current question, and the actual score obtained by the patient. These results are stored in the database. At the end of the test, both the maximum scores and the obtained scores for each question are summed separately to obtain a global score in the form $score/maxScore$ (e.g., a score of 18/22). This result can then be consulted by the patient. In order to follow the patient's progress, each time an evaluation test is performed, the platform sends an e-mail with a summary of the patient's performance to the therapist assigned to him/her.

Overall, these alterations contributed to building a simplified interface, suited for aged people, especially if cognitively impaired.

4. Automated tests

Since the selected neuropsychological tests comprise common or similar questions, we may approach its concrete implementation organized by type of question and the underlying technology with which they were implemented, rather than per test. Each type of question has set different challenges, each of which has been addressed individually with ad-hoc solutions. Overall, a total of 185 stimuli belonging to different types of tests have been selected for their implementation in the platform.

4.1. Naming objects and fingers

This type of stimuli belongs both to the MMSE and the ADAS-cog tests and evaluates a person's naming ability. Similarly to the exercises used in aphasia treatment, it consists of naming a series of objects that are shown in pictures, one at a time. These stimuli were implemented following a keyword spotting approach. A maximum score of 1 is given for each correct answer. The major innovation relative to the VITHEA exercises was the introduction of an optional semantic cue for some of the questions. This was implemented by adding a timer in the component responsible for the answer's recording and by making the virtual therapist to speak the cue after 20 seconds if no answer is detected. For this to be possible, both the clinician module and the internal structure of the database had to be extended for managing and storing the additional information. In fact, since the recording process is started from the beginning, the semantic cue is also recorded together with the patient's answer. Consequently, the logic of the patient module had also to be updated in order to remove the semantic cue spoken by the virtual therapist.

4.2. Repetition

The repetition question is part of the MMSE test and consists of repeating the following sentence: "*O rato roeu a rolha*" (the mouse gnawed the stopper). This question could be easily implemented with a keyword/key-phrase spotting approach, just like the ones for aphasia treatment. The maximum score is 1, which corresponds to a sentence correctly repeated.

4.3. Attention and calculation

This type of question belongs to the MMSE test, the idea is to successively subtract 3 beginning on 30 until 5 answers are given. In our first approach, we created a set of 5 different stimuli, each one asking separately for a specific calculation. These questions were also implemented with a keyword spotting approach. A score of 1 is given for each stimulus that corresponds to a correct answer, for a maximum score of 5.

4.4. Orientation to time, place and person

These type of stimuli are part both of the MMSE and the ADAS-cog, though some questions differ. They comprise stimuli intended to evaluate a person's orientation ability, asking the patient to report the current year, day, month, his name, the country and the town he lives in, among others. These are dynamic questions in the sense that there is not a universal answer to each question as it changes depending on the time, place and person. The solution was to provide several pre-compiled language models that were carefully structured so that, at any time, the platform knows which is the right model to chose. For the questions of orientation to person, the necessary information is acquired at the time of the creation of the user profile and then it is used to automatically generate the corresponding language models. The majority of these questions were implemented based on a standard keyword spotting approach. However, for the day and hour, it was necessary to create dedicated rule-based grammars. A correct answer is always scored with 1 point, while an incorrect answer scores 0.

4.5. Word recognition

The word recognition stimuli belong to the ADAS-Cog test and consist of presenting the patient a list with 12 words to learn, one at a time. Words are written in block letters on white cards. The learning process is made by asking the patient to read each word aloud and try to remember it. Then, a new list with 24 words is shown in the same way. This new list contains the 12 original words of the learning list, plus 12 new distracting words that are carefully chosen in terms of phonetic similarity and semantic meaning. For each word, the patient is then asked to indicate whether it was on the learning list or not. This whole process is repeated in 3 trials. Just like the day and hour questions, rule-based hand crafted recognition grammars for positive, negative or neutral answers were built. For the word recognition task itself, each presented word is individually scored. Specifically, a correct answer corresponds to a maximum score of 1, which yields a total score of 72 for the word recognition sub-test (i.e., 24 for each trial).

4.6. Evocation

Generally speaking, an evocation question consists of recalling a series of words, whether they have been previously learned or if they are subject to compliance with certain requirements. In either cases, the spoken answers produced for this kind of stimuli are commonly followed by filled pauses, i.e. hesitation sounds. For this reason, we adopted a keyword spotting approach that incorporates an ad-hoc model to deal with filled pauses. In terms of score, the calculation is processed by considering the number of keywords that are correctly produced, without repetitions.

For MMSE, the evocation question consists of the immediate and delayed recall of 3 words. This was implemented with an auditory stimuli for the immediate recall task and with a textual stimuli for the delayed recall task. The presentation of the evocation question that belongs to the ADAS-cog is very similar to the word recognition question. Basically, it consists of the immediate recall of a list with 10 words that were previously learned, the whole process is repeated in 3 trials.

4.7. Verbal Fluency

The animal naming question, which belongs to the Verbal Fluency test, is the most challenging among the evocation tests. This is explained by the fact that, contrarily to the other cases, which are based on a limited domain vocabulary tasks, this question comprises a more extended domain composed of the name of all known species of animals. Theoretically, the language model should cover all the known species of animals, however, in practice this is infeasible. Moreover, the size of the language model directly impacts the output of the ASR system. In fact, while a shorter list will cause the missing keywords to never be recognized, a longer list will increase instead the perplexity of the task.

The automatic creation of an ad-hoc language model for this type of question is detailed in [14], and is briefly reported here. The starting point consisted of an existing list of animal names [15] that included 6044 animal names, grouped, classified, and labelled with its semantic category, without inflected forms. Since some names are more likely and common than others, the initial list was used to build a probabilistic language model that exploited this information. The likelihood of each term was computed considering the total number of results that is returned by querying a web search engine. The retrieval strategy had to be refined several times in order to find the optimal approach, in fact initial queries have led to incorrect counts due to homonyms of some terms. The final approach consisted in using the animal name and the semantic category associated. Finally, the likelihood associated with each term also allows to sort the list numerically and thus to reduce its size by filtering out less popular terms. After several experiments, the language model that achieved the best results contained the 802 most popular animal names.

5. Experimental set-up

5.1. ASR/KWS system

The monitoring tool integrates the in-house ASR engine named AUDIMUS [16, 17], a hybrid recognizer that follows the connectionist approach [18]. The baseline system combines three MLP-based acoustic models trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), log-RelAtive SpecTrAl features (RASTA, 13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). These model networks were trained with 57 hours of downsampled Broadcast News data and 58 hours of mixed fixed-telephone and mobile-telephone data in European Portuguese [19]. The number of context input frames is 13 for the PLP and RASTA networks and 15 for the MSG network. Neural networks are composed by two hidden layers of 1500 units each one. Monophone units are modelled, which results in MLP networks of 39 soft-max outputs (38 phonemes + 1 silence) [12].

In order to support a keyword spotting approach, the baseline ASR system was modified to incorporate a competing background speech model that is estimated without the need for acoustic model re-training. In fact, while keyword models are described by their sequence of phonetic units provided by an automatic grapheme-to-phoneme module, the problem of background speech modelling must be specifically addressed. Here, the posterior probability of a background speech unit is estimated as the mean probability of the top-6 most likely outputs of the phonetic network at each time frame. In this way, there is no need for acoustic network re-training.

5.2. Portuguese cognitive impaired speech corpus

To evaluate the feasibility of the monitoring tool, we collected an ad-hoc speech corpus. This includes recordings of 5 people diagnosed with cognitive impairments and 5 healthy control subjects. All the participants are Portuguese native speakers.

Recordings took place in different environments with different acoustic conditions. In fact, healthy subjects were recorded in a quiet, domestic environment, while patients were recorded at CHPL, the Psychiatric Hospital of Lisbon. No particular constraints were imposed over background noise conditions. Each session consisted approximately of a 20 to 30-minutes recording. The data was originally captured with the platform at 16 kHz, and later down-sampled to 8 kHz to match the acoustic models sampling frequency. The collection of the patients data, besides being emotionally demanding, it is a valuable resource which implied logistic difficulties.

Table 1: Speech corpus data, including gender, age, education and diagnosis. B.E.: Basic Education, S.E.: Secondary Education, MCI: Mild Cognitive Impairment, AD: Alzheimer Disease, PTD: Post-traumatic Dementia

User	Gender	Age	Education	Diagnosis
1	M	86	B.E. - 1st Cycle	MCI
2	F	71	B.E. - 1st Cycle	AD
3	M	60	B.E. - 1st Cycle	PTD
4	F	79	Illiterate	AD
5	M	80	S. E.	MCI
6	F	67	B.E. - 1st Cycle	Healthy
7	F	72	B.E. - 1st Cycle	Healthy
8	M	76	B.E. - 1st Cycle	Healthy
9	F	74	B.E. - 1st Cycle	Healthy
10	M	76	B.E. - 1st Cycle	Healthy

5.3. Evaluation results

Due to the extensiveness of the ADAS-cog test, it was infeasible to evaluate all the implemented neuropsychological tests. In fact, we estimated that the total duration of the evaluation would have been more than two hours, which was considered unacceptable. Thus, only a representative subset of all the tests has been selected, comprising a total of 41 stimuli. We have considered different individual evaluation metrics, depending on the type of automated tests and a global evaluation focused on the targeted final application.

5.3.1. Evaluation of KWS-based tests

The Word Verification Rate (WVR) was used to assess the performance of the automatic evaluation module in the tests based on keyword spotting (KWS). This metric provides a measure of the reliability of the platform as a verification tool. In order to compute it, both manual and automatic transcriptions are processed to indicate, for each utterance, if the expected keyword has been said or not. Then, the WVR is computed for each speaker as the number of coincidences between the manual and automatic result (either true acceptances or true rejections) divided by the total number of exercises. Thus, a result closer to 1 is desirable. Table 2 presents the WVR computed for each speaker on all the tasks based on keyword spotting. Results are provided separately for those tests that rely on simple KWS (word-lists) and those based on rule-based grammars with competing background model (i.e.: hours, date, yes/no, etc.). In general, we can consider these results quite promising. In fact, they are comparable to those reported in [13], in an evaluation with aphasia patients. In this case, the average verification rates are considerably better with healthy users, which was expected due to the more challenging characteristics of the patients' data. Nevertheless, the performance achieved with cognitive impaired users is still quite promising. On the other hand, no significant differences can be observed regarding the KWS strategy (word-list vs. rule-based grammar).

Table 2: WVR by speaker for keyword spotting exercises.

Patients

User	KWS (word-list)	KWS (rule-based)
1	0.78	0.79
2	0.78	0.93
3	0.74	0.71
4	0.91	0.50
5	0.65	0.79
Avg. WVR	0.77	0.74

Healthy

User	KWS (word-list)	KWS (rule-based)
6	0.91	0.86
7	0.91	0.93
8	0.91	0.93
9	0.87	0.86
10	0.83	0.86
Avg. WVR	0.89	0.89

5.3.2. *Evaluation of evocation tests*

The evocation exercises differ from the keyword spotting exercises in the sense that the answers are not evaluated as right or wrong, but instead the number of terms correctly recalled is counted. For this reason, we started by evaluating the Word Error Rate (WER) between the reference (manual) and the hypothesis (automatic) users' answer. Evocation exercises are divided into two categories: they may contain a limited number of words to recall or, contrarily, they may consider an open domain of possible answers complying to a specific semantic domain (e.g., Animal Naming test). Thus, the evaluation was processed separately for the two categories. The average WER computed for patients and control group in the class of evocation exercises with a closed domain was 20.00% and 8.16%, respectively. However, the average WER computed for patients and control group on the Animal Naming test was much higher, 65.12% and 46.48%, respectively. After a closer analysis, we noticed that the substitutions were the main source of error. This may be explained by the poor language model used in this type of question, since this is based on an extensive list of unigrams. Basically, the size of the list impacts greatly the performance of the ASR system by increasing its perplexity. Moreover, the list comprises uncommon animal names and some of them are quite short, which implies that even a small sound may be detected as an animal. It is interesting to notice, however, that although the results in terms of WER are clearly unsatisfactory, and demand further research, the number of animals recognized is not so different from the reference number of animals actually said.

5.3.3. *Global evaluation*

An evaluation analysis of the automatic tests closer to the final targeted application is necessary to better assess their possible applicability as part of an automatic screening platform. For a sub-set of the tests, a straightforward evaluation method consists of comparing the total manual and the automatic scores achieved by the user according to the scoring values described in section 4 for each type of test. In particular, the Mean Absolute Error (MAE) and the Mean Relative Absolute Error (MRAE) is used to measure the differences between the overall scores computed manually and automatically. The scores were calculated according to the traditional assessment that is performed when applying a neuropsychological test. Table 3 reports the MAE and MRAE for the previously reported subsets of stimuli and the maximum possible score for each test set, which corresponds to the maximum error achievable, in addition to the results for two specific screening tests: the MMSE and the Animal Naming test. For these two tests, the scores achieved by each speaker are also shown in Figures 2 and 3.

Table 3: MAE and MRAE (in brackets) by type of question and by neuropsychological test.

Question type / Test	Max. Score	Patients	Healthy
KWS (word-list)	23	3.00 (26%)	2.60 (12%)
KWS (rule-based)	14	2.80 (37%)	1.60 (15%)
Evocation (w/o animals)	11	0.80 (23%)	0.80 (11%)
MMSE	22	2.20 (21%)	2.80 (14%)
Animal Naming	∞	2.60 (24%)	1.80 (17%)

In general, the achieved results were better for healthy people than for patients. This is an expected result due to the impaired condition of the patients, which are reflected on the quality and coherence of their speech. The most common symptoms, even in less impaired subjects, are a reduced intensity, a reduced pitch, and a hoarse voice. Besides, quite often during the evaluation, patients started talking of general topics of their interest not related with the question under evaluation. It was also the case that sometimes the subject uttered his answer when the virtual therapist was still explaining the stimulus, thus resulting in overlapped speech. Finally, differently from healthy subjects, patients sometimes changed their mind while they were answering a question. This may increase the perplexity of the ASR result, especially when dealing with rule-based keyword spotting questions, due to the added complexity of the language models.

The MAE error reported for question type and test ranges from 0.80 to 3.00 for the patients, and from 0.80 to 2.80 for the control group. In relative terms, the mean relative error with respect to the manual scores (MRAE) ranges from 21% to 37% for the patients group, and from 12% to 17% for the control group. Notice that, since the scores achieved by healthy users are generally higher and since there are few differences between the two groups in terms of MAE, the MRAE for patients is considerably higher. Alternatively, it is worth comparing the MAE with the maximum possible score. This value depends on the number of stimuli selected for each test. For the Animal Naming test, the maximum score could not be computed since the number of elements a subject is able to name in the given time is unknown. In general, it can be observed that the difference between the automatic and the manual evaluation is relatively small compared to the maximum score. For instance, we can observe that the MAE for the questions based on keyword spotting is 3.00 out of 23 for the patients, which corresponds to 13%, and 2.60 out of 23 for control subjects, which corresponds to 11.3%. Overall, we consider these results a quite good performance, suggesting that the platform may be useful and reliable as a monitoring tool.

5.4. Discussion about the platform

The conducted evaluation and data collection also allowed us to collect important feedback about the platform itself. In fact, we

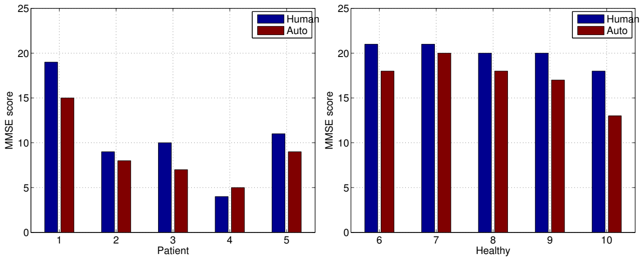


Figure 2: On the left side, MMSE scores of the human and automatic evaluation for the patient speakers. On the right side, MMSE scores of the human and automatic evaluations for the healthy speakers.

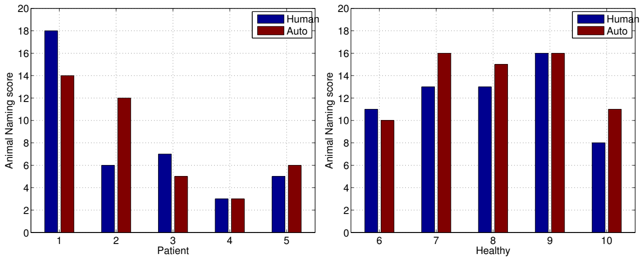


Figure 3: On the left side, Animal Naming scores of the human and automatic evaluation for the patient speakers. On the right side, Animal Naming scores of the human and automatic evaluations for the healthy speakers.

noticed that an advanced impaired condition may render more difficult the use of the system, specially when combined with deafness and with computer illiteracy, two factors that are associated with ageing. Patients with a more pronounced cognitive impairment or with auditory impairments, may have difficulties in understanding the question being asked. The computer illiteracy, however, may no longer be a problem in the not so distant future. Nevertheless, we expect that this tool will have adhesion for its usefulness and relevance. In fact, during test application, both the patients and healthy people demonstrated their appreciation for the platform, indicating that this is an interesting and appealing system. Moreover, they showed their interest in repeating the tests and using the platform regularly. Particularly, some of the patients were captivated by the animated virtual character, they liked its cartoon nature and the fact that it interacted with them verbally. This factor, together with the flexibility of the platform, let us think that in a near future the platform could be successfully turned in an environment useful both for training and monitoring cognitive skills. In fact, the kind of exercises that were adapted in the current version of the platform could be easily extended to the kind of games that are useful for stimulating brain activity, such as attention, memory etc. Further, the platform also allows to store recordings and evaluation results of each patient and make them available in an organized way, which can be useful for later consultation and comparison both by patients and by clinicians.

Finally, this platform raises interesting questions of ethical nature, i.e. whether such an automated tool should directly provide patients with a diagnosis similar to the one given by a clinician, or whether medical diagnosis should rather be produced exclusively by human doctors. One key related question is that diagnosis of mental disorders should always keep into account also normative data related with the language and education level of the patient. While we envision the possibility to incorporate the evaluations of such factors in future versions of the platform, these are not currently encompassed by our system. Finally, another important ethical question is whether patients should *always* be presented with the results of the automated tests. One may indeed argue that, in particular in presence of negative outcomes, the sensitivity of patients may be hurt and that, in such situations, it may be advisable to avoid exposing directly the tests' results to the patient and contact, instead, his/her relatives.

6. Conclusions and Future Work

In this work we developed an automatic web-based tool with SLT integration which could be used for monitoring cognitive impairments. The platform automates a set of neuropsychological tests that are commonly applied by therapists to assess the cognitive condition of a person. As far as we know, it is the only platform of this type implemented for the Portuguese population. The system has been assessed both with healthy subjects and patients. The mean absolute error between the manual and the automatic evaluation was relatively small, showing the feasibility of such type of system. We believe that this platform could be helpful for therapists and patients in the diagnosis of the disease. Its flexibility also allows the very easy creation of new exercises of the same type, with different stimuli. Besides, it could be easily extended to include different types of exercises that can be used for the daily training of cognitive abilities. For these reasons, we think that this tool could be an added value for society, helping in the prevention and in the early diagnosis of AD and mild cognitive impairments.

As future work we wish to remove completely the mouse interaction with the platform during the test application, automatically detecting when to stop the recording through silence detection technique. As the test currently already advances on its own when the recording is stopped, implementing this modification would enable to perform a complete test without any interaction, in a more agile way. Also, we plan to address the verbal fluency task, for which the preliminary results of the baseline system show much room for improvement. Finally, as mentioned in Section 5.4, medical diagnosis of dementia should keep into account also normative data related with the language and education level of the patient. In this sense, one open research question is how to automate the evaluation of these factors and incorporate them in the final diagnosis emitted by the system. Further, it would be desirable to extend the platform to incorporate intelligent filters aimed at identifying critical/negative outcomes, whose disclosure to the patient may risk hurting his/her sensitivity.

7. Acknowledgements

The authors of this work want to express their deep gratitude to Dr. José Salgado, clinical director of the Psychiatric Hospital of Lisbon Julio de Matos, for having made this work possible and for having provided us constant and full support. We also want to thank Dr. Inês Cunha and Dr. Vitorina Passão for their kind availability and precious feedback. This work was supported by national funds through – Fundação para a Ciência e a Tecnologia (FCT), under Grants SFRH/BD/97187/2013 and with reference UID/CEC/50021/2013.

8. References

[1] B. Nunes, "A demência em números," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[2] P. Moreira and C. Oliveira, "Fisiopatologia da doença de alzheimer e de outras demências," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[3] I. Santana, "A doença de alzheimer e outras demências - diagnstico diferencial," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[4] J. Barreto, "Os sinais da doença e a sua evolução," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[5] R. Coulston, E. Klabbers, J. Villiers, and J. Hosom, "Application of speech technology in a home based assessment kiosk for early detection of alzheimer's disease," in *Proc. Interspeech*, 2007.

[6] M. Guerreiro, "Avaliação neuropsicolgica das doenças degenerativas," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[7] E. Strauss, E. Sherman, and O. Spreen, *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*, 3rd ed. Oxford University Press, 2006.

[8] J. A. Anguera, J. Boccanfuso, J. L. Rintoul, O. Al-Hashimi, F. Faraji, J. Janowich, E. Kong, Y. Larraburo, C. Rolle, E. Johnston, and A. Gazzaley, "Video game training enhances cognitive control in older adults," *Nature*, vol. 501, pp. 97–101, 2013.

[9] M. Lehr, I. Shafran, and B. Roark, "Fully automated neuropsychological assessment for detecting mild cognitive impairment," in *In Interspeech*, 2012.

[10] S. S. Wang, P. D, J. Starren, and P. D, "A java speech implementation of the mini mental status exam."

[11] C. Amorim, "Automatic tool for screening of cognitive impairments," Master's thesis, Instituto Superior Tcnico, June 2014.

[12] A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins, "Automatic word naming recognition for an on-line aphasia treatment system," *Computer Speech & Language*, vol. 27, no. 6, pp. 1235 – 1248, 2013, special Issue on Speech and Language Processing for Assistive Technology.

[13] A. Abad, A. Pompili, A. Costa, and I. Trancoso, "Automatic word naming recognition for treatment and assessment of aphasia," in *Proc. Interspeech*, 2012.

[14] H. Moniz, A. Pompili, F. Batista, I. Trancoso, A. Abad, and C. Amorim, "Automatic recognition of prosodic patterns in semantic verbal fluency tests - an animal naming task for edutainment applications," in *18TH INTERNATIONAL CONGRESS OF PHONETIC SCIENCES*, 2015.

[15] N. J. Mamede, J. Baptista, C. Diniz, and V. Cabarrao, "String: An hybrid statistical and rule-based natural lan- guage processing chain for portuguese," in *International Conference on Computational Processing of Portuguese Propor*, 2012.

[16] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.Media: a Broadcast News speech recognition system for the European Portuguese language," in *Proc. International Conference on Computational Processing of Portuguese Language (PROPOR)*, 2003.

[17] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "The L^2F Broadcast News Speech Recognition System," in *Proc. Fala2010*, 2010.

[18] N. Morgan and H. Bourlad, "An introduction to hybrid HMM/connectionist continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[19] A. Abad and J. Neto, "Automatic classification and transcription of telephone speech in radio broadcast data," in *Proc.International Conference on Computational Processing of Portuguese Language (PROPOR)*, 2008.

Extending a Dutch Text-to-Pictograph Converter to English and Spanish

Leen Sevens, Vincent Vandeghinste, Ineke Schuurman, Frank Van Eynde

Centre for Computational Linguistics
KU Leuven, Belgium

firstname@ccl.kuleuven.be

Abstract

We describe how a Dutch Text-to-Pictograph translation system, designed to augment written text for people with Intellectual or Developmental Disabilities (IDD), was adapted in order to be usable for English and Spanish. The original system has a language-independent design. As far as the *textual* part is concerned, it is adaptable to all natural languages for which interlingual WordNet [1] links, lemmatizers and part-of-speech taggers are available. As far as the *pictographic* part is concerned, it can be modified for various pictographic languages. The evaluations show that our results are in line with the performance of the original Dutch system. Text-to-Pictograph translation has a wide application potential in the domain of Augmentative and Alternative Communication (AAC). The system will be released as an open source product.

Index Terms: Augmentative and Alternative Communication, Pictographic Languages, Text-to-Pictograph Translation

1. Introduction

In our daily lives, we are constantly confronted with pictographs. Think of traffic signs, signs in buildings that direct visitors to the elevators, the meeting rooms, the toilets, and the emergency exits, or signs for telling people that dogs need to be kept on a leash (see Figure 1).

Figure 1: Pictographs in our daily lives.

Similar pictographs are used as a form of Augmentative and Alternative Communication (AAC). AAC assists people with severe communication disabilities to be more socially active in interpersonal interaction, learning, education, community activities, employment, volunteering, and care management. Schools, institutions, and sheltered workshops use specific pictographs that are related to everyday activities and objects to allow accessible written communication between children or adults with Intellectual or Developmental Disabilities (IDD) and their caregivers, in an offline setting.

It is undeniable that current technological advances influence our lives in various aspects. Not being able to access or use information technology is a major form of exclusion. In order to reduce social isolation, there is an acute need for digital picture-based communication interfaces that enable contact for people with IDD. Adding pictographs to text can provide help in reading and understanding the text. It is estimated that between two and five million people in the European Union could benefit from symbols or symbol-related text as a means of written communication [2].

The Dutch Text-to-Pictograph translation system that is described in Vandeghinste et al. [3] is used in the WAI-NOT[1] communication platform. WAI-NOT is a Flemish, non-profit organization that gives people with severe communication disabilities the opportunity to familiarize themselves with computers, the internet, and social media. The website makes use of an email client that automatically augments written text with a series of Beta[2] or Sclera[3] pictographs. WAI-NOT's first translation system would rely on a simple one-on-one match between the input words and the pictograph file names, usually leading to erroneous translations and leaving many words untranslated. Vandeghinste et al. [3] improved this engine by introducing linguistic analysis. Their Text-to-Pictograph translation system was made as language-independent as possible.

Within the framework of Able to Include,[4] which aims to improve the living conditions of people with IDD, we built English and Spanish versions of this system. English and Spanish being a Germanic and a Romance language, respectively, we show that the engine manages to generalize well over different European language families.

After a discussion of related work (section 2), we introduce the Beta and Sclera pictograph sets (section 3), followed by an explanation of how existing links between WordNets can be used to automatically connect pictographs to words in source languages other than Dutch (section 4). In the remainder of this paper, we describe the system's general architecture (section 5). The evaluations (section 6) show that our results are in line with the performance of the Dutch system. Section 7 shows that the Text-to-Pictograph system has a wide application potential in the domain of AAC. Finally, we describe our conclusions and future work (section 8).

2. Related work

Pictographic communication has grown from local initiatives, some of which have scaled up to larger communities. Across Europe, many pictograph sets are in place, such as Blissymbolics,[5] PCS,[6] Pictogram,[7], ARASAAC,[8] Widgit,[9] Beta, and Sclera.

[1] http://www.wai-not.be/

[2] https://www.betasymbols.com/

[3] http://www.sclera.be/

[4] http://abletoinclude.eu

[5] http://blissymbolics.org/

[6] http://www.mayer-johnson.com/category/symbols-and-photos

[7] http://www.pictogram.se/

[8] http://www.catedu.es/arasaac/

[9] https://widgit.com/

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 110–117,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

Many of the problems that written languages encounter can be overcome by the use of pictographic languages. For instance, they can be understood across language barriers[10] [4] and there is less ambiguity involved. Pictographic communication systems for remote, online communication include Messenger Visual, an instant messaging service [5], Communicator [6], Pictograph Chat Communicator III [7], and VIL, a Visual Inter Lingua [4]. Mihalcea and Leong [8] argue that the understanding of graphical sentences is similar to that of target language texts obtained by means of machine translation. Leemans [4] shows that an appropriately designed iconic language, built according to a set of fixed principles, leads to no difference in the recognition rate of icons for people of western and non-western culture, yielding an average rate of about 79%. None of these abovementioned authors, however, consider users with IDD when designing the system.

Other pictograph-based communication systems are specifically designed for people with IDD. Patel et al. [9] introduce Image-Oriented Communication Aid, an interface using the Widgit symbol set, allowing users to build picture-supported messages on a touch screen computer. Motocos [10] are image exchange devices that are designed for children with autism, including audio cues for easier understanding of the image cards. The mobile application PhotoTalk [11] aids people with aphasia by providing a digital photograph management system in support of verbal communication. Nevertheless, all these systems require face-to-face communication in an offline setting.

The use of online information technology systems as a way to enhance the quality of life of people with IDD is a recent development. For accessible, remote communication, Keskinen et al. [2] introduce SymbolChat, a platform for picture-based instant messaging, where the interaction is based on touch screen input and speech output. The Text-to-Pictograph conversion system described in Vandeghinste et al. [3] applies shallow linguistic analysis to Dutch input text and automatically generates sequences of Beta and Sclera pictographs, allowing people with IDD to read messages independently. Only few other publications related to the task of translating texts for pictograph-supported communication can be found in the literature, such as Goldberg et al. [12] and Mihalcea and Leong [8], but these systems do not translate the whole sentence or are not focused on IDD.

3. Pictographic languages

Mihalcea and Leong [8] note that complex and abstract concepts (such as *democracy*) are not always easy to depict. Some characteristics of natural languages may not be present in the pictographic languages.[11] Usually, no distinction between singular and plural is made. Tense, aspect, and inflection information is removed, and so are the auxiliaries and the articles.[12] Pictographic languages are simplified languages, that are often specifically designed for people with IDD.

Although experiments with the Pictogram set [13] have revealed that many pictographs are difficult and wrongly interpreted, a correct interpretation is easily accepted and remembered without any problem. By giving people with speech and language disorders the opportunity to familiarize themselves with the pictographs, they learn to interpret the symbols more easily. However, a deliberate effort is needed.

The Text-to-Pictograph translation system currently gives access to two pictograph sets, Sclera and Beta (see Figure 2).

Sclera pictographs[13] are mainly black-and-white pictographs, although colour is sometimes used to indicate permission (green) or prohibition (red). Over 13,000 pictographs are available and more are added upon user request. Sclera pictographs often represent *complex* concepts, such as a verb and its object (such as *to feed the dog*) or compound words (such as *carrot soup*). There are hardly any pictographs for adverbs or prepositions.

The *Beta* set[14] consists of more than 3,000 coloured pictographs. Easy recognition being one of the main objectives, Beta is characterized by its overall consistency and the use of different types of arrows and dashes (pointing to an object, indicating changes in space or time or depicting actions). Beta hardly contains any complex pictographs. Most of the pictographs represent *simplex* concepts.

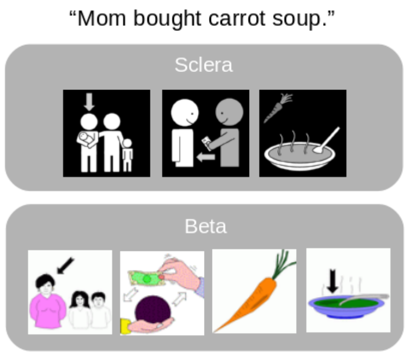


Figure 2: Example of a sentence being translated into Sclera and Beta pictographs. Tense information is removed. The Sclera translation contains a complex pictograph, namely *carrot soup*.

4. Linking pictographs to other WordNets

WordNets, lexical-semantic databases, are an essential component of the Text-to-Pictograph translation system. For the original Dutch system, Cornetto [14, 15] was used. Its English and Spanish counterparts are Princeton WordNet 3.0 [1][15] and the Spanish Multilingual Central Repository (MCR) 3.0 [16].[16] WordNets contain synsets (groupings of synonyms that have an abstract, usually numeric identifier, see Figure 3) and are designed in such way that each synset is connected to one or more lemmas.

Vandeghinste and Schuurman [17] manually linked 5710 Sclera pictographs and 2760 Beta pictographs to Dutch synsets in Cornetto.[17] An essential step in building Text-to-Pictograph translation systems for other languages is making sure that the pictographs are connected to (sets of) words in those languages.

[10] Although cultural differences remain.

[11] We use the term *pictographic language* in order to refer to the combination of individual pictographs, that belong to a specific *pictograph set*, into a larger meaningful structure.

[12] There are some exceptions. Beta, for instance, contains the Dutch articles.

[13] Freely available under Creative Commons License 2.0.

[14] The coloured pictographs can be obtained at reasonable prices, while their black-and-white equivalents are available for free.

[15] http://wordnet.princeton.edu/

[16] http://adimen.si.ehu.es/web/MCR/

[17] As a Cornetto license can no longer be obtained, the authors will transfer these links to the Open Source Dutch WordNet (http://wordpress.let.vupr.nl/odwn/).

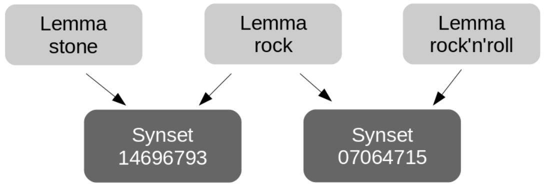


Figure 3: An example of a lemma, *rock*, having different meanings and belonging to different synsets. Two synsets are shown here.

Manually linking thousands of pictographs all over again would be a very time-consuming procedure. Instead, by transferring the connections automatically (see Figure 4), this process can be sped up drastically.

Sevens et al. [18] note that connections between WordNets are an important resource in knowledge-based multilingual language processing. The already mentioned Cornetto database for Dutch, used to build the Dutch Text-to-Pictograph translation system, contains connections to the English Princeton WordNet. We describe how we automatically connected Beta and Sclera pictographs to synsets in Princeton WordNet 3.0 in section 4.1.

Many WordNets nowadays contain high-quality links between the source language's synsets and Princeton WordNet 3.0, which is often viewed as the *central* WordNet. Princeton WordNet 3.0 now also plays this central role in our Text-to-Pictograph translation system. Having obtained the links between Beta and Sclera pictographs and Princeton WordNet 3.0, it becomes possible to automatically assign pictographs to synsets in any WordNet that has decent connections with Princeton WordNet,[18] allowing us to quickly build Text-to-Pictograph translation systems for many other languages. For example, with the English pictograph connections in place, a mapping between the pictographs and Spanish synsets in MCR 3.0 became possible. This process is described in section 4.2.

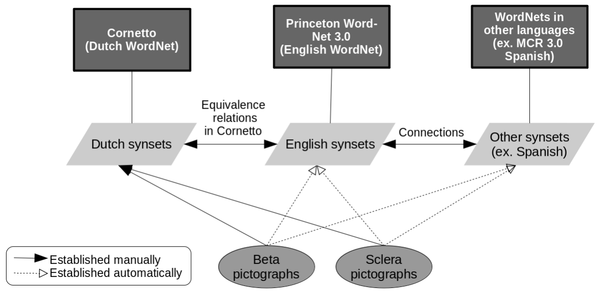


Figure 4: Making Princeton WordNet 3.0 the central WordNet of the Text-to-Pictograph translation system and transferring the links to the MCR 3.0 for Spanish.

4.1. Connecting pictographs to Princeton WordNet 3.0

Cornetto's *equivalence relations* establish connections between Dutch and English synsets in Princeton WordNet. These relations have originally been established semi-automatically by Vossen et al. [19], filling the database with more than 80000 links between Dutch and English synsets.

Sevens et al. [18] showed that a considerable amount of the original links were highly erroneous, making them not yet very reliable for multilingual processing. By using these equivalence relations, we would risk assigning pictographs to unrelated synsets in Princeton WordNet 3.0. In the case of a Dutch synset being wrongly connected to an English synset, writing a message in English would allow the system to generate pictographs that depict another concept. Therefore, we used the filtered,[19] more reliable connections that were established by Sevens et al. [18].

As a result, it became possible to automatically assign a large amount of Sclera and Beta pictographs to English synsets in Princeton WordNet 3.0. However, 154 (5.58%) Beta pictographs and 288 (5.04%) Sclera pictographs still had to be connected manually, either because the original equivalence relation was rejected by the filtering algorithm, or because the Dutch compound word corresponded to multiple words in English and forced us to treat the pictograph as a complex pictograph[20] in English (such as the Dutch word *vanillesuiker*, meaning *vanilla sugar* in English). In some rare cases, no equivalent English concept existed in the WordNet for an existing Dutch concept (for instance, the fictional character *Zwarte Piet* or typical kinds of food such as *choco*, which can roughly be translated as *chocolate spread*).

4.2. Connecting pictographs to the Spanish MCR 3.0

The MCR 3.0 integrates in the same EuroWordNet framework WordNets from five different languages, namely English, Catalan, Spanish, Basque, and Galician. Words in one language are connected to words in any of the other languages through Inter-Lingual-Indexes. Sevens et al. [18] showed that the links between English and Spanish synsets were correctly established, making it possible for us to create highly reliable connections between Beta and Sclera pictographs and Spanish synsets. This exact same process can be done for any language's WordNet that establishes reliable links to Princeton WordNet 3.0.

5. The Text-to-Pictograph translation system for English and Spanish

In this section, we describe how a textual message is converted into a sequence of Sclera or Beta pictographs [3] (see Figure 5), with an application to English and Spanish. The source text first undergoes shallow linguistic analysis (section 5.1). For further processing, two routes can be taken. The semantic route is only applied to content words (nouns, verbs, adjectives, adverbs) that are present in the WordNets. It consists of linking the source text to synsets in the databases (section 5.2) and retrieving the pictographs that are connected to these synsets (section 5.3). The direct route (section 5.4), which runs in parallel with the semantic route, contains specific rules for appropriately dealing with pronouns, and it uses a dictionary for parts-of-speech that are not present in the WordNets. The system contains a handful of parameters (section 5.5), which were tuned beforehand (section 5.6). Finally, as explained in section 5.7, an optimal sequence of pictographs is selected.

[18] A full list can be found on http://globalwordnet.org/wordnets-in-the-world/

[19] Filtering was done by using large bilingual dictionaries.

[20] A pictograph that is connected to multiple synsets instead of just one synset. For example, the pictograph depicting vanilla sugar is connected to both the synset that contains the lemma *vanilla* and the synset that contains the lemma *sugar*.

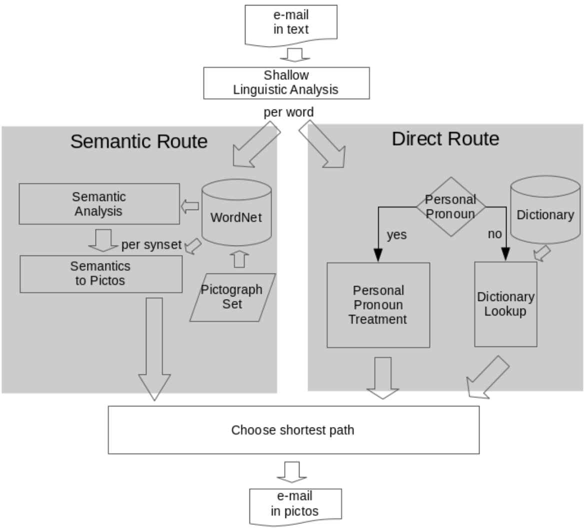
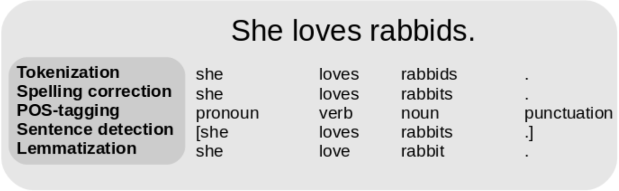


Figure 5: Architecture of the translation engine.

5.1. Shallow linguistic analysis

The source text undergoes shallow linguistic processing, consisting of several sub-processes (see Figure 6). This process is analogous to the linguistic processing step in the original Dutch tool.


She loves rabbids.

Tokenization	she	loves	rabbids	.
Spelling correction	she	loves	rabbits	.
POS-tagging	pronoun	verb	noun	punctuation
Sentence detection	[she	loves	rabbits	.]
Lemmatization	she	love	rabbit	.

Figure 6: An example of shallow linguistic processing.

First, *tokenization* is applied to split the punctuation signs from the words, with the exception of the hyphen/dash and the apostrophe, as they often belong to the word.

As the targeted users have different levels of illiterateness, basic *spelling correction* (one deletion, one insertion, one substitution)[21] aids in finding the correct variant of words that do not appear in the lexicon[22] and the list of first names.[23]

Next, *part-of-speech tagging* is applied. For English, we used HunPos [20], an open source tagger, using the English training data (with Penn Treebank tags[24]) made available on its website.[25] For Spanish, part-of-speech tagging (with TreeTagger tags[26]) and lemmatization are done in one step with TreeTagger [21].[27] TreeTagger is available for a large variety of European languages.

The Text-to-Pictograph translation system works on the sentence level. Although most messages sent by the users only contain one sentence, *sentence detection* is applied. Segmentation is based on full stops, which will eventually correspond to line breaks in the resulting pictographic representation.

The next step is *lemmatization*, which requires a language-specific treatment. For English, we built a lemmatizer based on a list of English token/part-of-speech combinations and their lemma.[28] As mentioned before, for Spanish, part-of-speech tagging and lemmatization are done with TreeTagger.

One additional adaptation concerns the treatment of the Spanish *pro-drop* phenomenon (which occurs in all Romance languages, with the exception of French), meaning that personal pronouns in subject position are usually omitted (unless emphasis is given). Translating such a message into pictographs would leave us with no subject, as the pictographic representations of words are based on the lemma form and do not retain any grammatical information. However, person information can be inferred from the verb in the source sentence. We wrote a set of rules that explicitly adds the personal pronouns in the message before converting it into a series of pictographs.[29] When a matching personal pronoun is already found within a window of three words (since adverbs or pronouns can appear between the subject and the verb), these rules are not applied (see Figure 7).

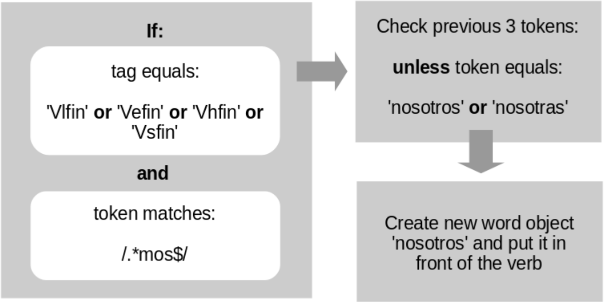


Figure 7: An example of a pro-drop rule. The tags correspond to *finite lexical verb*, *finite estar (to be)*, *finite haber (to have)*, and *finite ser (to be)*. The token has to end on *-mos*, which indicates a first person plural form. *Nosotros* and *nosotras* correspond to the English pronoun *we*.

5.2. Semantic analysis

The first step in the semantic analysis consists of the detection of words with a negative polarity, such as *not/no* and *no/ningún*. When such a word is found, the system looks for its head (a verb or a noun) and adds the value *negative* to its polarity feature.

For each word in the source text, the system returns all possible WordNet synsets (see section 4). The synsets are filtered,

[21] We are currently designing a spelling corrector that is specifically tailored towards Dutch text written by people with IDD. Our approach does not rely on the use of parallel corpora (erroneous text - corrected text). Therefore, it can also prove to be useful for spelling correction in other languages.

[22] http://www.anc.org/SecondRelease/frequency2.html (for English) and http://corpus.leeds.ac.uk/frqc/internet-es-forms.num (for Spanish)

[23] http://www.quietaffiliate.com/free-first-name-and-last-name-databases-csv-and-sql (for English and Spanish)

[24] http://www.cis.upenn.edu/ treebank

[25] https://code.google.com/p/hunpos/downloads/list

[26] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/data/spanish-tagset.txt

[27] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[28] http://www.anc.org/

[29] When the verb is a third person singular or plural, these rules are not applied, as its subject could be a noun phrase. This problem can be solved by applying deeper grammatical analysis. Gender information (*he*, *she*, *it*), however, cannot be inferred from the verb alone and requires deeper semantic knowledge.

keeping only those where the part-of-speech tag of the synset matches the part-of-speech tag of the word.

Certain links between lemmas and synsets can be disabled in order to remove unwanted, often sexual meanings of common words, which are not appropriate for some groups of users (such as one meaning of the word *member*).

5.3. Retrieving the pictographs

The WordNet synsets described in section 5.2 are used to connect pictographs to natural language text. This way, the lexical coverage of the system is greatly improved, as pictographs are connected to sets of words that have the same meaning, instead of just individual words. Additionally, if a synset is not covered by a pictograph, the links between synsets can be used to look for alternative pictographs with a similar meaning. For instance, the *hyperonymy* relation can be used if no pictograph is found for a concept that is too specific (such as *rabbit* for *cottontail*, see Figure 8). The *antonymy* relation, indicating that synsets are the opposite of each other, selects a pictograph of the antonym, along with a negation pictograph (such as *not sick* for *recovered*). The *XPos* relation concerns similar words with a different part-of-speech tag (such as the adjective *female* for *woman*). However, using pictographs through *synset propagation* (making use of the WordNet relations) is controlled by parameters or penalties for not using the proper concept (see section 5.5).

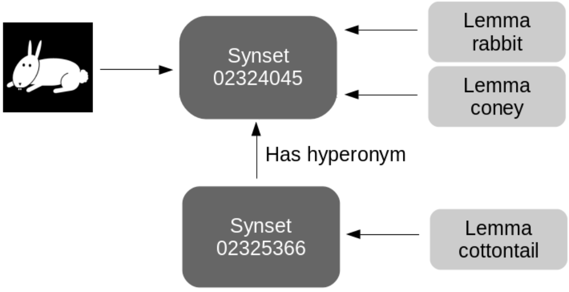


Figure 8: When a specific word, such as *cottontail*, does not have a pictograph connected to its synset, WordNet relations will be used to find a similar concept and display its pictograph instead. The synset for *rabbit* and *coney* (a synonym of *rabbit*) is found.

5.4. The direct route

The English and Spanish WordNets contain nouns, verbs, adjectives, and adverbs. To deal with pronouns and words that have a part-of-speech tag that is not covered by the WordNets, the direct route is introduced.

To make sure that *personal and possessive pronouns* are covered, they are given an explicit treatment. Person, gender, and number information can be obtained during the part-of-speech tagging process, resulting in correct pictograph translations.

The *dictionary* provides a direct link between the *token/lemma/tag* and the names of the pictographs. The *tag* field and either the *lemma* or *token* field can be left underspecified. For instance, in Sclera, there is a direct link between the lemma *hey* and the pictograph *hallo-zeggen-2.png* (*to say hello*), while the verb *miss* needs an additional *verb* tag to avoid confusion with the noun. The dictionary is used to cover any words that are missing from the database, because their part-of-speech tag is not included in the WordNet database (such as various types of greetings), or because the concept is too recent (such as *tablet*), among other things.

5.5. The parameters

For every word in the sentence, the system checks whether one or more pictographs can be found for it and whether the use of these pictographs is subject to a penalty. Penalties correspond to parameters that were tuned beforehand.

The first set of parameters (*hyperonym* penalty, *antonym* penalty, and *XPos* penalty) concern the maximum distance (*threshold* parameter) allowed between the original text and the pictographic message in terms of synset relations (see section 5.3).

The second set of parameters is related to the *numeric features* of the pictographs (*no number* and *wrong number*), as some pictographs make a distinction between singular or plural concepts (such as *oog.png*, depicting one eye, and *ogen.png*, depicting two eyes).

The last set of parameters determines the behaviour as to *the route to take*. An *out-of-vocabulary* parameter penalizes for leaving a content word untranslated, while the *direct route* parameter is a negative penalty (i.e. a bonus) for choosing the direct route over the semantic route.

Furthermore, the use of complex pictographs, which reunite multiple concepts within one pictograph (see section 3), will be preferred by the system over the separation of those concepts. The shorter the pictographic translation is, the higher it will be scored by the system (see section 5.7).

5.6. Tuning the parameters

The parameters that are mentioned in section 5.5 are tuned for every natural language/pictographic language pair. Ideally, tuning would be based on emails or text messages written by people with IDD. These messages are usually short, tend to refer to everyday life and very often contain spelling mistakes, like tweets.[30] As we did not have a large corpus of messages written by the targeted users at our disposition, we selected 75 English tweets and 75 Spanish tweets based on the following criteria: the messages should contain at least 8 words, they have to refer to personal experiences (no citations or lyrics), and they are allowed to contain spelling mistakes or lack punctuation marks. The tweets were retrieved by searching for messages containing the hash tags *#school/#escuela*, *#love/#amor*, *#family/#familia*, *#happy/#feliz*, and *#sad/#triste*.

For both languages, we manually translated, to the best of our ability, all tweets into Beta and Sclera pictographs. We built a local hill climber that varies the parameters (see section 5.5) when running the Text2Picto script on each of the four test sets (from English and Spanish to Beta and Sclera). The BLEU metric [22] was used as an indicator of relative improvement. In order to maximize the BLEU score, we ran five trials of a local hill climbing algorithm for each natural language/pictographic language pair. We did this until BLEU converged onto a fixed score after several thousands of iterations. Each trial was run with random initialization values, while varying the parameters between certain boundaries and with a granularity (size of the parameter steps) of one in order to cover different areas of the search space. From these trials, we took the best scoring parameter values for all four language/pictographic language pairs.

[30] https://twitter.com/

		With proper names		Without proper names	
Condition	**Precision**	**Recall**	**F-Score**	**Recall**	**F-Score**
Sclera					
Baseline	71.37%	61.25%	65.92%	62.25%	66.50%
Add frequent concepts	93.30%	71.95%	81.25%	73.04%	81.94%
Rel. improv.	*30.73%*	*17.47%*	*23.26%*	*17.33%*	*23.22%*
Beta					
Baseline	75.08%	70.63%	72.78%	71.71%	73.36%
Add frequent concepts	82.56%	85.07%	83.80%	86.14%	84.31%
Rel.improv.	*9.96%*	*20.45%*	*15.14%*	*20.12%*	*14.93%*

Table 1: Manual evaluation of the English system

		With proper names		Without proper names	
Condition	**Precision**	**Recall**	**F-Score**	**Recall**	**F-Score**
Sclera					
Baseline	73.84%	57.63%	64.74%	58.30%	65.16%
Add frequent concepts	93.31%	82.17%	87.38%	83.14%	87.93%
Rel. improv.	*26.37%*	*42.58%*	*34.97%*	*42.61%*	*34.95%*
Beta					
Baseline	83.48%	60.83%	70.38%	61.26%	70.66%
Add frequent concepts	94.64%	86.01%	90.12%	86.83%	90.57%
Rel.improv.	*13.37%*	*41.39%*	*28.05%*	*41.74%*	*28.18%*

Table 2: Manual evaluation of the Spanish system

5.7. Selecting the optimal path

An A* algorithm[31] calculates the optimal pictographic sequence for the source text. Its input is the pictographically annotated source message, together with the pictographs' penalties, depending on the number and kind of synset relations the system had to go through to connect them to the words.

The algorithm starts with a queue containing an empty path that still has all the input words left to process. In every step, the currently best scoring pictograph path is extended. We check whether there are any pictographs, with their corresponding penalties, connected to the next word that has to be processed.[32] New paths are thus created by adding the retrieved pictograph to the list of the already matched pictographs. All possible paths are added to the queue. The queue is sorted by lowest estimated cost and the best scoring path is extended. This process is repeated until the first queue element no longer has any words left to process.

When encountering words that have their *antonym* feature set to *negative* (see section 5.2), we insert the negation pictograph.

6. Evaluation

At the time of our evaluation, we did not yet have a corpus of messages written by people at IDD at our disposition. An evaluation set was built using the selection procedure as described in section 5.6. A total of 50 English tweets and 50 Spanish tweets were retrieved.

After having obtained the system's output translations for every message from the evaluation set, we performed a manual verification with one judge, who removed untranslated non-content words (such as *just*, *although*, and *it* in English). This allowed calculating the recall. For each of the translated words, she judged whether the pictograph generated was the correct pictograph, in order to calculate precision. As proper names occur rather frequently in online environments, we have calculated recall and F-score with and without proper names, in the latter case removing all proper names from the output. Precision remains the same in both conditions. In the case where proper names are included, they are not converted into pictographs, affecting recall negatively. In applications, similar to an option that is currently available in the WAI-NOT environment, proper names occurring in the contact lists of the users can be converted into the photographs that are attached to user profiles, resulting in more personalized messages.

Using the automatic pictograph connections that Sevens et al. [18] created by using the links between Cornetto synsets and Princeton WordNet synsets and the links between Princeton WordNet synsets and Spanish MCR synsets, a baseline system could be built. This system, which is not subject to any post-editing actions in the WordNet databases, leaves us with F-Scores of 66.50% and 73.36% for Sclera and Beta, respectively, for English text without proper names. For Spanish, F-Scores of 65.16% and 70.66% are obtained. A decent baseline system was thus created by making use of the previously available WordNet relations.

To improve the English and Spanish systems, we added or edited the 500 most frequently used words according to the Dutch WAI-NOT corpus,[33] in order to cover the specific vocabulary that the target group uses to address their peers or caregivers. For each one of these words, we translated them into English and Spanish and checked whether the right pictograph was connected to its synset. If this was not the case, we disabled the erroneous pictographs or created new pictograph connections. Sometimes, the pictograph dictionary (direct route) was used to add missing words to the database, such as different types of greetings. As a result, the English system currently yields F-Scores of 81.94% and 84.31% for Sclera and Beta, respectively, while the Spanish system reaches F-Scores of 87.93% and 90.57%, both for text in which proper names are omitted.

These results are comparable to the manual evaluations for Dutch [3]. The authors obtain F-Scores of 87.16% and 87.27% for Sclera and Beta translations of Dutch IDD text, respectively.

7. Application potential

The Text-to-Pictograph translation system will be released as an open source product, allowing developers to build pictograph-supported AAC applications and web browser extensions.

The pictographs are not meant to replace written text. They can be used as a stepping stone towards a better comprehension of written content.

Since textual content on the web, in particular long or difficult words, is sometimes very challenging for the target group to deal with, a Text-to-Pictograph translation system in the form of a web browser extension could be a welcome addition for many users. Web browser extensions are programs that extend the functionality of a web browser. For instance, by hovering over a difficult word, the program could show the pictographic representation of that word. This idea has already been implemented by the creators of Widgit, although their Point system[34] does not make use of semantic networks to simplify extension to additional languages.[35]

The system offers the possibility for family members, caregivers, and teachers to build pictographic messages more easily. Browsing large databases to find the appropriate icons is a long and tedious job, that can be facilitated by automatically translating a textual message into a series of pictographs. This

[31] A pathfinding algorithm that uses a heuristic to search the most likely paths first.

[32] If a complex pictograph is retrieved, the system checks whether the other synsets that belong to that complex pictograph are connected to any of the remaining words to process. If this is the case, the word that is linked to that synset is removed from the list of words to process.

[33] A corpus containing more than 40000 e-mails sent by users with IDD and their caregivers. Most e-mails are about their everyday life.

[34] https://widgit.com/products/online.htm

[35] We thank the anonymous reviewers for this observation.

way, pictograph-supported instructions, schedules and menus will become easier to construct. Text-to-Pictograph translation will also allow the family members and caregivers to send pictographic e-mails to the target group, making it simpler to communicate in an online setting, where the use of written text would normally cause big difficulties.

Within the Able to Include framework, a mobile app is currently being developed to address a variety of scenarios in which pictographs offer support. The tool will also integrate text-to-speech and text simplification technologies. The user can choose a technology (or a combination of technologies, such as text simplification followed by translation into pictographs) that he or she feels most comfortable with.

While our system is initially focused on users with IDD (since the tool was developed on the request of WAI-NOT, a website for people with disabilities), its general architecture can be reused in various other contexts, such as education, language learning for non-native speakers, and translation into sign languages.

8. Conclusions and future work

We have shown how the Dutch Text-to-Pictograph translation system can be extended towards other languages. To implement new languages, only a few components are required: decent connections between the source language's WordNet and the Princeton WordNet 3.0 (as we have shown for Spanish), a language-specific part-of-speech tagger and lemmatizer, a new set of parameters to optimize the system's performance and possibly some additional rules to deal with language-specific properties.

Future work will consist of improving the English and Spanish systems. Proper word sense disambiguation will have to be applied, as the system currently only takes the most frequent sense for a given word. We will look into possibilities for better spelling correction, specifically tailored towards text written by people with cognitive disabilities, and simplification of the pictographic output. Finally, the inverse relation, pictograph-to-text translation, will also be taken care of, allowing users to create textual messages by selecting a series of pictographs [23].

In collaboration with Faculty of Psychology and Educational Sciences of KU Leuven and our Able to Include partners, the pictograph translation system will be tested by the target group. The results will give us better insights concerning the usability of the engine.

Analysis of text written by English and Spanish users with IDD will reveal which concepts are missing from the databases and we will continue to improve the coverage of the system.

9. References

[1] G. Miller, R. Beckwidth, C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: An On-line Lexical Database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[2] T. Keskinen, T. Heimonen, M. Turunen, J. Rajaniemi, and S. Kauppinen, "SymbolChat: A Flexible Picture-based Communication Platform for Users with Intellectual Disabilities," *Interacting with Computers*, vol. 24, no. 5, pp. 374–386, 2012.

[3] V. Vandeghinste, I. Schuurman, L. Sevens, and F. Van Eynde, "Translating Text into Pictographs," *Natural Language Engineering*, Accepted.

[4] P. Leemans, *VIL: A Visual Inter Lingua*. Dissertation. Worcester Polytechnic Institute., 2001.

[5] P. Tuset, J. Barbern, P. Cervell-Pastor, and C. Janer, "Designing Messenger Visual, an Instant Messenging Service for Individuals with Cognitive Disability," in *IWAAL 1995 – Proceedings of 3rd International Workshop on Ambient Assisted Living*, 1995, pp. 57–64.

[6] T. Takasaki and Y. Mori, "Design and Development of a Pictogram Communication System for Children around the World," in *IWIC 2007 – Proceedings of the 1st International Conference on Intercultural Collaboration*, 2011, pp. 193–206.

[7] J. Munemori, T. Fukada, M. Yatid, T. Nishide, and J. Itou, "Pictograph Chat Communicator III: a Chat System that Embodies Cross-Cultural Communication," in *KES 2010 – Proceedings of the 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems: Part III*, 2012, pp. 473–482.

[8] R. Mihalcea and C. Leong, "Toward Communicating Simple Sentences Using Pictorial Representations," *Machine Translation*, vol. 22, no. 3, pp. 153–173, 2009.

[9] R. Patel, S. Pilato, and D. Roy, "Beyond Linear Syntax: an Image-Oriented Communication Aid," *ACM Journal of Assistive Technology: Outcomes and Benefits*, vol. 1, no. 1, pp. 57–66, 2004.

[10] G. Hayes, S. Hirano, G. Marcu, M. Monibi, D. Nguyen, and M. Yeganyan, "Interactive Visual Supports for Children with Autism," *Personal Ubiquitous Computing*, vol. 14, no. 1, pp. 663–680, 2010.

[11] M. Allen, J. McGrenere, and B. Purves, "The Field Evaluation of a Mobile Digital Image Communication Application Designed for People with Aphasia," *ACM Transactions on Accessible Computing*, vol. 1, no. 1, 2008.

[12] A. Goldberg, X. Zhu, C. Dyer, M. Eldawy, and L. Heng, "Easy as ABC? Facilitating Pictorial Communication via Semantically Enhanced Layout," in *CoNLL 2008 – Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 2008.

[13] K. Falck, *The Practical Application of Pictogram*, Lycksele, 2001.

[14] P. Vossen, I. Maks, R. Segers, and H. van der Vliet, "Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database," in *LREC 2008 – Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008.

[15] H. van der Vliet, I. Maks, P. Vossen, and R. Segers, "The Cornetto Database: Semantic issues in Linking Lexical Units and Synsets," in *EURALEX 2010 – Proceedings of the 14th EURALEX 2010 International Congress*, 2010.

[16] A. G. Agirre, E. Laparra, and G. Rigau, "Multilingual Central Repository Version 3.0: Upgrading a Very Large Lexical Knowledge Base," in *LREC 2012 – Proceedings of the 8th International Conference on Language Resources and Evaluation*, 2012.

[17] V. Vandeghinste and I. Schuurman, "Linking Pictographs to Synsets: Sclera2Cornetto," in *LREC 2012 – Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014, pp. 3404–3410.

[18] L. Sevens, V. Vandeghinste, and F. Van Eynde, "Improving the Precision of Synset Links Between Cornetto and Princeton WordNet," in *LG-LP 2014 – Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing*, 2014.

[19] P. Vossen, L. Bloksma, and P. Boersma, *The Dutch Wordnet. EuroWordNet Paper*, Amsterdam, 1999.

[20] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos - an Open Source Trigram Tagger," in *ACL 2007 – Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume, Proceedings of the Demo and Poster Sessions*, 2007, pp. 209–212.

[21] H. Schmid, "Improvements in Part-of-speech Tagging with an Application to German," in *SIGDAT 1995 – Proceedings of the ACL SIGDAT-Workshop*, 1995.

[22] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Evaluation of Machine Translation,” in *ACL 2002 – Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[23] L. Sevens, V. Vandeghinste, I. Schuurman, and F. Van Eynde, “Natural Language Generation from Pictographs,” in *ENLG 2015 – Proceedings of the 15th European Workshop on Natural Language Generation*, 2015.

Individuality-Preserving Spectrum Modification for Articulation Disorders Using Phone Selective Synthesis

Reina Ueda, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe, 6578501, Japan
{reina_1102, aihara}@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

This paper presents a speech synthesis method for people with articulation disorders resulting from athetoid cerebral palsy. For people with articulation disorders, there are duration, pitch and spectral problems that cause their speech to be less intelligible and make communication difficult. In order to deal with these problems, this paper describes a Hidden Markov Model (HMM)-based text-to-speech synthesis approach that preserves the voice individuality of those with articulation disorders and aids them in their communication. For the unstable pitch problem, we use the F0 patterns of a physically unimpaired person, with the average F0 being converted to the target F0 in advance. Because the spectrum of people with articulation disorders is often unstable and unclear, we modify generated spectral parameters from the HMM synthesis system by using a physically unimpaired person's spectral model while preserving the individuality of the person with an articulation disorder. Through experimental evaluations, we have confirmed that the proposed method successfully synthesizes intelligible speech while maintaining the target speaker's individuality.

Index Terms: Articulation disorders, Speech synthesis system, Hidden Markov Model, Assistive Technologies

1. Introduction

In this study, we focus on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. About two babies in 1,000 are born with cerebral palsy [1]. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. It is classified into the following types: 1) spastic, 2) athetoid, 3) ataxic, 4) atonic, 5) rigid, and a mixture of these types [2]. Athetoid symptoms develop in about 10-15% of cerebral palsy sufferers [1]. In the case of persons with articulation disorders resulting from the athetoid type of cerebral palsy, his/her movements are sometimes more unstable than usual. That means their utterances (especially their consonants) are often unstable or unclear due to their athetoid symptoms, and there is a great need for voice systems that can assist them in their communication.

An HMM-based speech synthesis system [3] is a text-to-speech (TTS) system that can generate signals from input text data. A TTS system may be useful for those with articulation disorders because they have difficulty moving their lips. In an HMM-based speech synthesis system, the spectrum, F0 and duration are modeled simultaneously in a unified framework. Mel-cepstral coefficients are used as spectral features, which are modeled by continuous density HMMs. F0 patterns are modeled by a hidden Markov model based on multi-space probability distribution (MSD-HMM [4]), and state duration densities are modeled by single Gaussian distributions [5].

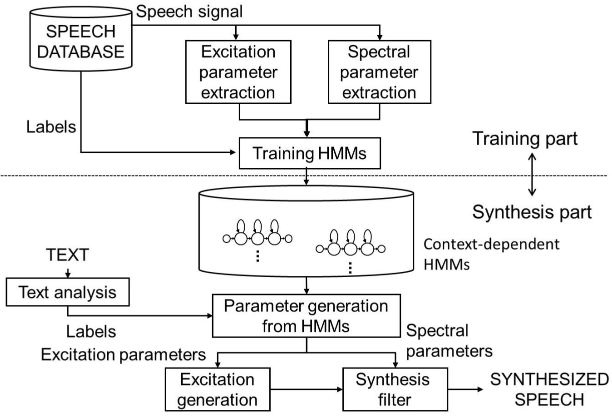


Figure 1: HMM-based sound synthesis system

In the field of assistive technology, Veaux *et al.* [6] used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting from Amyotrophic Lateral Sclerosis (ALS). They have proposed a reconstruction method for degenerative speech disorders using an HMM sound synthesis system. In this method, the subject's utterances are used to adapt an average voice model pre-trained on many speakers. Creer *et al.* [7] also adapt the average voice model of multiple speakers to the severe dysarthria data. And Khan *et al.* [8] uses such adaption method to the laryngectomy patient's data. Yamagishi *et al.* [9] proposed a project called "Voice Banking and Reconstruction". In that project, various types of voices were collected, and they proposed TTS for ALS using that database. Also, Rudzicz [10] proposed a speech adjustment method for people with articulation disorders based on observations from the database.

In this paper, we propose an HMM-based speech synthesis method for articulation disorders because there are several problems in the recorded voice of persons with articulation disorders, and this causes the output synthesized signals to be unintelligible. To deal with these problems, it is necessary to develop a speech synthesis system in which the output signals become more intelligible and include the subject's individuality.

To generate an intelligible voice while preserving the speaker's individuality, we train the speech synthesis system using training data from both a person with an articulation disorder and a physically unimpaired person. Because the utterance rate of persons with articulation disorders differs from that of a

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 118–123,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

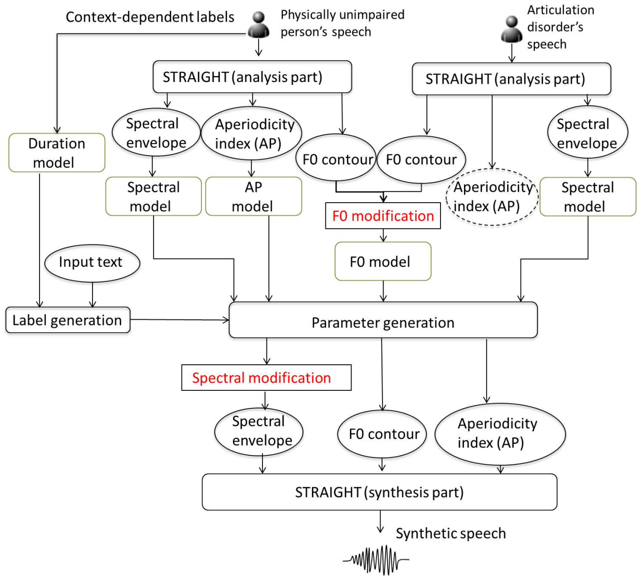


Figure 2: Diagram of HMM-based sound synthesis method for articulation disorders

physically unimpaired person, we utilize the duration model of a physically unimpaired person only in our method. In addition to the utterance rate problem, the F0 patterns of persons with articulation disorders are often unstable compared to those of physically unimpaired persons. In our method, the F0 model is trained from a physically unimpaired person's F0 patterns, and the average F0 is used as the F0 pattern for the person with an articulation disorder.

As for the spectral problem associated with persons with articulation disorders, the consonant parts of their speech are often unstable or unclear, which causes their voice to be unintelligible. To resolve this consonant problem, we conduct different operations on the consonant and vowel parts. For the consonants parts, we basically generate the output spectrum from the spectral model of a physically unimpaired person. For the vowel parts, we generate the output spectrum from the spectral model of a person with an articulation disorder in order to preserve the person's individuality.

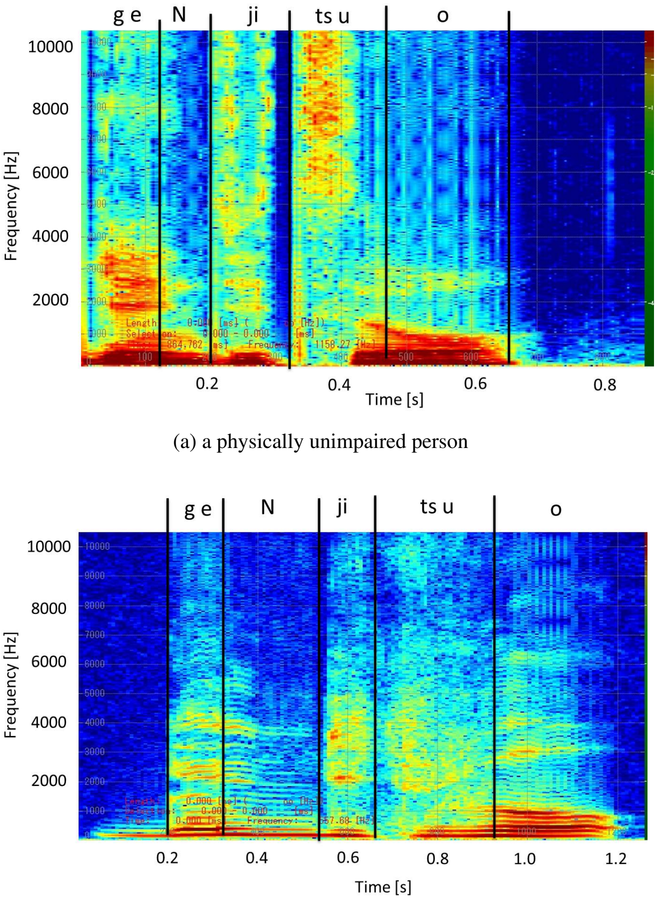
2. HMM-based sound synthesis

2.1. Basic approach

Fig. 1 shows the overview of the basic approach to text-to-speech synthesis (TTS) based on HMMs. This figure shows the training and synthesis parts of the HMM-based TTS system. In the training part, parameters (spectral, F0, and aperiodicity) are extracted as feature vectors. These features are modeled by context-dependent HMMs. Also, by installing the duration model, it is able to model each parameter, as well as the duration in the unified framework.

In the synthesis part, a context-dependent label sequence is obtained from an input text by text analysis. A sentence HMM is constructed by concatenating context-dependent HMMs according to the context-dependent label sequence. Then, HMM state sequences $q = [q_1, \cdots, q_T]$ are decided from the duration model as follows:

$$\hat{q} = \arg\max_{q} P(q|\lambda) \tag{1}$$



(a) a physically unimpaired person

(b) a person with an articulation disorder

Figure 3: Examples of spectrogram uttered for // g e N j i ts u o

where T, q_t, and λ represent the number of frames, index of the HMM-state of the t-th input frame, and the parameter sets of HMM, respectively. The explicit constraint between static and dynamic features, and signal parameter sets are generated with maximizing HMM likelihood [11].

$$c = \arg\max_{c} P(\mathbf{W}c|\hat{q}, \lambda) \tag{2}$$

In Eq. (2), $c = [{c_1}^\top, \cdots, c_t^\top, \cdots, c_T^\top]^\top$ represents signal parameter sequences, $c_t = [c(1), \cdots, c(D)]^\top$ represents a signal parameter vector of the t-th frame, and $\mathbf{W}$ represents the matrix constructed from weights which are used for calculating dynamic features [12].

Finally, by using an MLSA (Mel-Log Spectrum Approximation) filter [13], speech is synthesized from the generated parameters.

2.2. HMM-based sound synthesis for articulation disorders

If each feature parameter is trained using the acoustic features obtained from a person with an articulation disorder, the synthesized sound becomes unintelligible. Therefore, we created a more intelligible synthesized sound while preserving the speaker's individuality by mixing the voices of a person with an articulation disorder and a physically unimpaired person.

Fig. 2 shows the overview of our method. In this method, we train the speech synthesis system using training data from both a person with an articulation disorder and a physically unimpaired person. First, we extract three acoustic parameters (F0 contour, spectral envelope, and aperiodicity index (AP)) from these two person's speaking voices by using STRAIGHT analysis [14]. After extracting the features, the F0 patterns of a physically unimpaired person are modified as explained in Section 2.3.

Because the duration of persons with articulation disorders is slower than that of physically unimpaired people, the duration model is generated using only the context-dependent label sequences of a physically unimpaired person. With the input text and the duration model, context-dependent label sequences are generated. Then, spectral, F0 and AP parameters are generated based on the label sequences and trained HMMs, where F0 parameters are generated from the modified F0 model and AP parameter sequences are generated from the AP model of a person with an articulation disorder.

Each spectral parameter is generated from each person's spectral model. After parameter generation, the spectral parameters of a person with an articulation disorder are modified as explained in Section 2.4. Finally, the output signal is synthesized from the features (spectral envelope, F0 contour, and aperiodicity index) by using the synthesis part of the STRAIGHT. In the following section, we explain the details of the operations related to spectral and F0 parameters.

2.3. F0 modification

In this method, the F0 patterns of a physically unimpaired person are used for training the F0 model in HMM synthesis because the F0 patterns of a person with an articulation disorder are often unstable. To make the F0 feature's characteristics close to those of a person with an articulation disorder, the F0 features of a physically unimpaired person are modified to those of a person with an articulation disorder. The F0 model is trained from the converted F0 sequences, which means that the F0 model includes the individuality of a person with articulation disorder.

The F0 features of a physically unimpaired person are modified by using the following linear transformation:

$$\hat{x}_t = \frac{\sigma_y}{\sigma_x}(x_t - \mu_x) + \mu_y \tag{3}$$

where x_t represents the log-scaled F0 of the physically unimpaired person at the frame t, μ_x and σ_x represent the mean and standard deviation of x_t, respectively. μ_y and σ_y represents the mean and standard deviation of the log-scaled F0 of a person with an articulation disorder, respectively.

2.4. Spectral modification

Fig. 3 shows the original spectrograms for the word "genjitsuo" ("real" in English) of a physically unimpaired person and a person with an articulation disorder. As shown in Fig. 3, the high-frequency spectral power of a person with an articulation disorder is weaker compared to that of a physically unimpaired person. This fact implies that the synthesized spectrum of the consonant components for a person with an articulation disorder becomes weak, which makes the person's speech difficult to understand.

For the spectral vowel components, the spectral parameters of a person with an articulation disorder are needed in order to preserve the target individuality. As shown in Fig. 2, after being given the input text, we generate spectral parameter sequences from each person's spectral model. Then, we create the combined spectral parameter sequences, which include the parameters of a physically unimpaired person at the high-frequency part and the parameters of a person with an articulation disorder at the low-frequency part. This combination of spectral parameters is given by

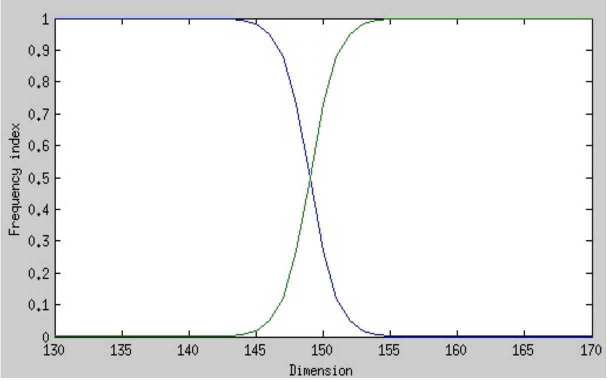


Figure 4: Plot of the function f_m and f_g (green: f_m blue: f_g)

$$\hat{S}^{(ij)} = f_m^{(j)} S_m^{(ij)} + f_g^{(j)} S_g^{(ij)} \tag{4}$$

where S_m, S_g, $\hat{S}$, i and j represent the spectrum of a physically unimpaired person, the spectrum of a person with an articulation disorder, the modified spectrum, the index of spectral frames, and the frequency index, respectively. The weight functions are given by

$$f_m^{(j)} = \frac{1}{1 + e^{(-j+S)}} \tag{5}$$

$$f_g^{(j)} = \frac{1}{1 + e^{(j-S)}} \tag{6}$$

where f_m represents the weight function for a physically unimpaired person's spectrum, f_g represents that of a person with an articulation disorder, and S represents the control parameter, respectively.

Fig. 4 shows an example of the functions f_m and f_g. The function, f_m, emphasizes the high-frequency components and weakens the low-frequency components of spectral parameters. The function, f_g, emphasizes the low frequency components and weakens the high-frequency components of spectral parameters.

By using Eq. (4), at the high-frequency part, the spectrum is complemented by that of a physically unimpaired person in order to make the consonants clear. At the low-frequency part, we need to preserve the spectrum of a person with an articulation disorder in order to preserve the individuality. The spectral modification is calculated at each frame using Eq. (4), and the frequency thresholds are determined for the vowel part and consonant part. In our study, the total number of spectral dimensions (indexes) is 513, S is set to 150 for the vowel part, and S is set to 80 for the consonant part.

3. Experiments

3.1. Experimental conditions

We prepared the training data for two men. One is a physically unimpaired person, and the other is a person with an articula-

Table 1: Voices compared in the evaluation tests

Type	Duration Model	F0 Model	AP Model	Spectral Model
ADM	AD	AD	AD	AD
Ref1	PU	AD	AD	AD
Prop	PU	convPU	AD	MIX
Ref2	PU	convPU	AD	AD
PUM	PU	PU	PU	PU

Note
ADM: Articulation disorder person's model
Prop: Proposed method
PUM: Physically unimpaired person's model
AD: Articulation Disordered
PU: Physically Unimpaired
convPU: Creating the model from a physically unimpaired person's pa
which are converted to those of the person with an articulati
MIX: mixing ADM and PUM spectra using Eq. (4)

tion disorder. We used 513 sentences from the ATR Japanese database for a physically unimpaired person, and recorded 429 sentences in the same database uttered by a person with an articulation disorder. The speech signals were sampled at 48 kHz and the frame shift was 5 ms. Acoustic and prosodic features were extracted by using STRAIGHT. As spectral parameters, mel-cepstrum coefficients, their dynamic, acceleration coefficients were used. As excitation parameters, log-F0 and 5 band-filtered aperiodicity measures [15] were used and their dynamic and acceleration coefficients were also used. Context-dependent phoneme HMMs with five states were used in the speech synthesis system [3].

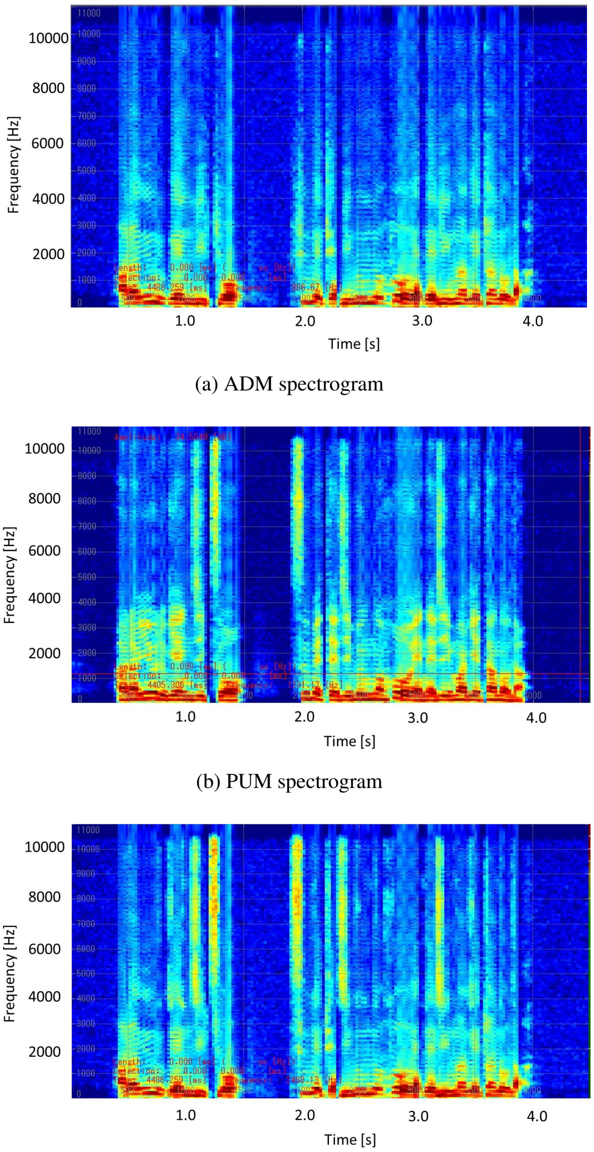
In order to confirm the effectiveness of our method, we evaluated both the aspect of listening intelligibility and the aspect of speaker similarity by listening to voices synthesized under the five conditions shown in Table 1. Ten sentences included in the ATR Japanese database were synthesized under those five conditions. A total of 8 Japanese speakers took part in the listening test using headphones. For speaker similarity, we performed a MOS (Mean Opinion Score) test [16]. In the MOS test, the opinion score was set to a 5-point scale (5: Identical, 4: Very Similar, 3: Quite Similar, 2: Dissimilar, 1: Very Dissimilar). For the listening intelligibility, a paired comparison test was carried out, where each subject listened to pairs of speech converted by the two methods, and then selected which sample was more intelligible.

3.2. Results and discussion

Table 2: Average duration per mora in 50 sentences

	Average time [ms/mora]
ADM	219.768
PUM	179.69

We calculated the average synthesized signal's duration per mora in 50 sentences. As shown in Table 2, the average duration of ADM (Articulation disorder person's model) is 219.768 [ms/mora] and that of PUM (Physically unimpaired person's model) is 179.69 [ms/mora]. As compared to the duration of PUM, that of ADM is quite slower, which causes the unintelligibility of the synthesized sound.



(a) ADM spectrogram

(b) PUM spectrogram

(c) Modified spectrogram

Figure 5: Examples of synthesized spectrograms

In the proposed method, we generated the modified spectral parameters by mixing both ADM and PUM spectral parameters. Fig. 5a shows the generated spectrum from the ADM spectral model and Fig. 5b shows the generated spectrum from the PUM spectral model. Both spectral parameters are generated from the same text and the same PUM duration model so that they have the same number of frames and dimensions. As shown in Fig. 5a, the high-frequency component is weaker compared to Fig. 5b, which means that the consonant parts of ADM spectral parameters are weak. This causes the output synthesized signals to be less intelligible. Fig. 5c shows the modified spectrum created from both ADM and PUM spectral parameters by using Eq. (4). As shown in Fig. 5c, the consonant parts are complemented by the high-frequency parameters of PUM while preserving ADM's low-frequency components.

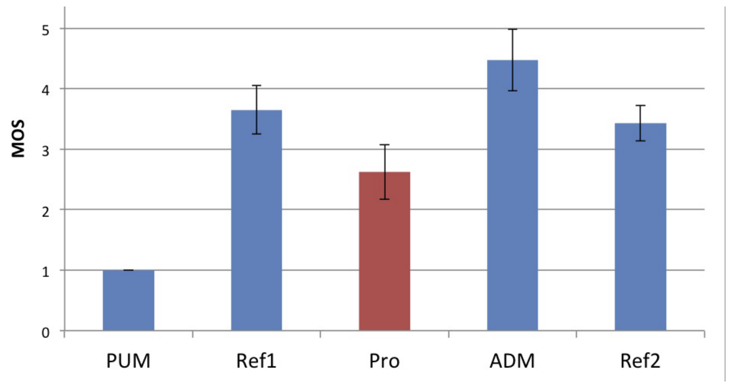


Figure 6: Speaker similarity to the articulation disorder person's speech

Fig. 6 shows the results of the MOS test on speaker similarity, where the error bar shows a 95% confidence score. As shown in Fig. 6, the ADM score was the highest score of all. This is because the signal from ADM is synthesized only from the feature parameters of a person with an articulation disorder. The Prop score is slightly less than those of Ref1 and Ref2 because of the modification of the spectral parameters.

Fig. 7 shows the preference score for the listening intelligibility, where the error bar shows a 95% confidence score. As shown in Fig. 7, our method obtained a higher score than Ref1 and ADM. These results show that the proposed method is effective. By replacing the physically unimpaired person's duration model and converting his F0 patterns to those of the person with an articulation disorder improves intelligibility. Our method also obtained a higher score than Ref2. This result shows that modifying the output spectral parameters is quite effective in improving intelligibility. Therefore, considering from Figs. 6 and 7, it is confirmed that our proposed method implements the synthesized signals which is intelligible and includes individuality of a person with an articulation disorder.

4. Conclusion

We have proposed a text-to-speech synthesis method based on HMMs for a person with an articulation disorder. In our method, to generate synthesized sounds that are more intelligible, the duration model of a physically unimpaired person is used, and the F0 model is trained using the F0 features of a physically unimpaired person, where the average F0 is converted to the articulation disorder person's F0 using a linear transformation. In order to complement the consonant parts of the spectrum of a person with an articulation disorder, we replaced the high-frequency parts with those of a physically unimpaired person. The experimental results showed that our method is highly effective in improving the listening intelligibility of speech spoken by a person with an articulation disorder. In future research, we will complement the consonant parts of the spectral parameters at the training part.

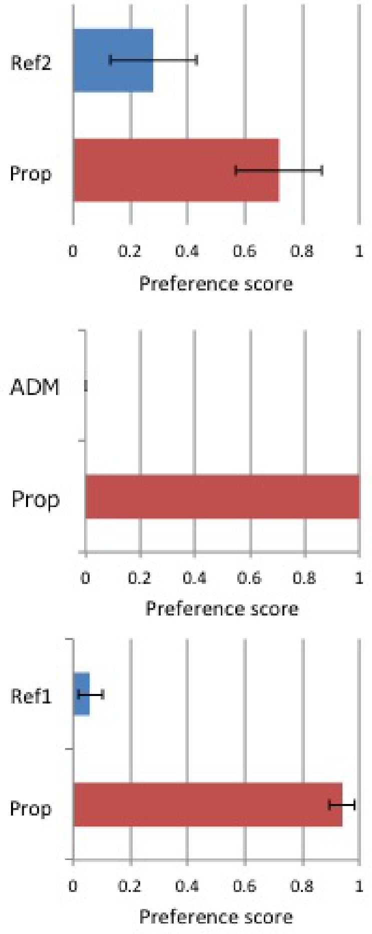


Figure 7: Preference scores for listening intelligibility

5. References

[1] M. V. Hollegaard, K. Skogstrand, P. Thorsen, B. Norgaard-Pedersen, D. M. Hougaard, and J. Grove, “Joint analysis of SNPs and proteins identifies regulatory IL18 gene variations decreasing the chance of spastic cerebral palsy,” *Human Mutation*, vol. 34, pp. 143–148, January 2013.

[2] T. Canale and W. C. Campbell, *Campbell’s operative orthopaedics*. Technical report, Mosby Year Book, June 2002, vol. 12.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Proc. of Eurospeech*, 1999, pp. 2347–2350.

[4] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling,” in *Proc. of ICASSP*, 1999, pp. 229–232.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration Modeling in HMM-based Speech Synthesis System,” in *Proc. of ICSLP*, 1998, pp. 29–32.

[6] C. Veaux, J. Yamagishi, and S. King, “Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders,” in *Proc. of Interspeech*, 2012.

[7] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, “Building personalised synthetic voices for individuals with severe speech impairment,” *Computer Speech & Language*, vol. 27, no. 6, pp. 1178–1193, 2013.

[8] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham, “Reconstructing the voice of an individual following laryngectomy,” *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–66, 2011.

[9] J. Yamagishi, C. Veaux, S. King, and S. Renals, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[10] F. Rudzicz, “Adjusting dysarthric speech signals to be more intelligible,” *Computer Speech and Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[11] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc. of ICASSP*, 2000, pp. 1315–1318.

[12] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[13] S. Imai, K. Sumita, and C. Furuichi, “Mel log spectrum approximation (MLSA) filter for speech synthesis,” *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, pp. 10–18, 1983.

[14] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds,” *Speech communication*, vol. 27, pp. 187–207, 1999.

[15] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT.” in *Proc. of MAVEBA*, 2001, pp. 59–64.

[16] I. T. Union, “ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology,” International Telecommunication Union, Tech. Rep., July 2006.

Recognition of Distress Calls in Distant Speech Setting: a Preliminary Experiment in a Smart Home

Michel Vacher[1], *Benjamin Lecouteux*[2], *Frédéric Aman*[1],
Solange Rossato[2], *François Portet*[2]

[1]CNRS, LIG, F-38000 Grenoble, France
[2]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
41 rue Mathématiques, BP 53, 38041 Grenoble cedex9, France

`Michel.Vacher@imag.fr, Benjamin.Lecouteux@imag.fr, Frederic.Aman@imag.fr,`
`Solange.Rossato@imag.fr, Francois.Portet@imag.fr`

Abstract

This paper presents a system to recognize distress speech in the home of seniors to provide reassurance and assistance. The system is aiming at being integrated into a larger system for Ambient Assisted Living (AAL) using only one microphone with a fix position in a non-intimate room. The paper presents the details of the automatic speech recognition system which must work under distant speech condition and with expressive speech. Moreover, privacy is ensured by running the decoding on-site and not on a remote server. Furthermore the system was biased to recognize only set of sentences defined after a user study. The system has been evaluated in a smart space reproducing a typical living room where 17 participants played scenarios including falls during which they uttered distress calls. The results showed a promising error rate of 29% while emphasizing the challenges of the task.

Index Terms: Smart home, Vocal distress call, Applications of speech technology for Ambient Assisted Living

1. Introduction

Life expectancy has increased in all countries of the European Union in the last decade. Therefore the part of the people who are at least 75 years old has strongly increased and solutions are needed to satisfy the wishes of elderly people to live as long as possible in their own homes. Ageing can cause functional limitations that –if not compensated by technical assistance or environmental management– lead to activity restriction [1][2]. Smart homes are a promising way to help elderly people to live independently at their own home, they are housings equipped with sensors and actuators [3][4][1][5]. Another aspect is the increasing risk of distress, among which falling is one of the main fear and lethal risk, but also blocking hip or fainting. The most common solution is the use of kinematic sensors worn by the person [6] but this imposes some constraints in the everyday life and worn sensors are not always a good solution because some persons can forget or refuse to wear it. Nowadays, one of the best suited interfaces is the voice-user interface (VUI), whose technology has reached maturity and is avoiding the use of worn sensors thanks to microphones set up in the home and allowing hands-free and distant interaction [7]. It was demonstrated that VUI is useful for system integrating speech commands [8].

The use of speech technologies in home environment requires to address particular challenges due to this specific environment [9]. There is a rising number of smart home projects considering speech processing in their design. They are related to wheelchair command [10], vocal command for people with dysarthria [11][8], companion robot [12], vocal control of appliances and devices [13]. Due to the experimental constraints, few systems were validated with real users in realistic situation condition like in the SWEET-HOME project [14] during which a dedicated voice based home automation system was able to drive a smart home thanks to vocal commands with typical people [15] and with elderly and visually impaired people [16].

In this paper we present an approach to provide assistance in a smart home for seniors in case of distress situation in which they can't move but can talk. The challenge is due to expressive speech which is different from standard speech: is it possible to use state of the art ASR techniques to recognize expressive speech? In our approach, we address the problem by using the microphone of a home automation and social system placed in the living room with ASR decoding and voice call matching. In this way, the user must be able to command the environment without having to wear a specific device for fall detection or for physical interaction (e.g., a remote control too far from the user when needed). Though microphones in a home is a real breach of privacy, by contrast to current smart-phones, we address the problem using an in-home ASR engine rather than a cloud based one (private conversations do not go outside the home). Moreover, the limited vocabulary ensures that only speech relevant to the command of the home is correctly decoded. Finally, another strength of the approach is to have been evaluated in realistic conditions. The paper is organised as follow. Section 2 presents the method for speech acquisition and recognition in the home. Section 3, presents the experimentation and the results which are discussed in Section 5.

2. Method

The distress call recognition is to be performed in the context of a smart home which is equipped with e-lio[1], a dedicated system for connecting elderly people with their relatives as shown in Figure 1. e-lio is equipped with one microphone for video conferencing. The typical setting and the distress situations were determined after a sociological study conducted by the GRePS laboratory [17] in which a representative set of seniors were included.

From this sociological study, it appears that this equipment

[1] http://www.technosens.fr/

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 124–129,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

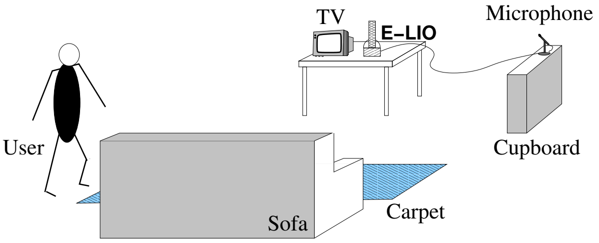


Figure 1: Microphone position in the smart home

is set on a table in the living room in font of the sofa. In this way, an alert could be given if the person falls due to the carpet or if it can't stand up from the sofa. This paper presents only the audio part of the study, for more details about the global audio and video system, the reader is referred to [18].

2.1. Speech analysis system

The audio processing was performed by the software CIRDOX[19] whose architecture is shown in Figure 2. The microphone stream is continuously acquired and sound events are detected on the fly by using a wavelet decomposition and an adaptive thresholding strategy [20]. Sound events are then classified as noise or speech and, in the latter case, sent to an ASR system. The result of the ASR is then sent to the last stage which is in charge of recognizing distress calls.

In this paper, we focus on the ASR system and present different strategies to improve the recognition rate of the calls. The remaining of this section presents the methods employed at the acoustic and decoding level.

2.2. Acoustic modeling

The Kaldi speech recognition tool-kit [21] was chosen as ASR system. Kaldi is an open-source state-of-the-art ASR system with a high number of tools and a strong support from the community. In the experiments, the acoustic models were context-dependent classical three-state left-right HMMs. Acoustic features were based on Mel-frequency cepstral coefficients, 13 MFCC-features coefficients were first extracted and then expanded with delta and double delta features and energy (40 features). Acoustic models were composed of 11,000 context-dependent states and 150,000 Gaussians. The state tying is performed using a decision tree based on a tree-clustering of the phones. In addition, off-line fMLLR linear transformation acoustic adaptation was performed.

The acoustic models were trained on 500 hours of transcribed French speech composed of the ESTER 1&2 (broadcast news and conversational speech recorded on the radio) and REPERE (TV news and talk-shows) challenges as well as from 7 hours of transcribed French speech of the SH corpus (SWEET-HOME) [22] which consists of records of 60 speakers interacting in the smart home and from 28 minutes of the Voix-détresse corpus [23] which is made of records of speakers eliciting a distress emotion.

2.2.1. Subspace GMM Acoustic Modelling

The GMM and Subspace GMM (SGMM) both model emission probability of each HMM state with a Gaussian mixture model, but in the SGMM approach, the Gaussian means and the mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections.

The SGMM model [24] is described in the following equations:

$$\begin{cases} p(\mathbf{x}|j) = \sum\limits_{m=1}^{M_j} c_{jm} \sum\limits_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i), \\ \mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \\ w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}. \end{cases}$$

where $\mathbf{x}$ denotes the feature vector, $j \in \{1..J\}$ is the HMM state, i is the Gaussian index, m is the substate and c_{jm} is the substate weight. Each state j is associated to a vector $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S is the phonetic subspace dimension) which derives the means, μ_{jmi} and mixture weights, w_{jmi} and it has a shared number of Gaussians, I. The phonetic subspace $\mathbf{M}_i$, weight projections $\mathbf{w}_i^T$ and covariance matrices $\mathbf{\Sigma}_i$ i.e; the globally shared parameters $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$ are common across all states. These parameters can be shared and estimated over multiple record conditions.

A generic mixture of I gaussians, denoted as Universal Background Model (UBM), models all the speech training data for the initialization of the SGMM.

Our experiments aims at obtaining SGMM shared parameters using both SWEET-HOME data (7h), Voix-détresse (28mn) and clean data (ESTER+REPERE 500h). Regarding the GMM part, the three training data set are just merged in a single one. [24] showed that the model is also effective with large amounts of training data. Therefore, three UBMs were trained respectively on SWEET-HOME data, Voix-détresse and clean data. These tree UBMs contained 1K gaussians and were merged into a single one mixed down to 1K gaussian (closest Gaussians pairs were merged [25]). The aim is to bias specifically the acoustic model with the smart home and expressive speech conditions.

2.3. Recognition of distress calls

The recognition of distress calls consists in computing the phonetic distance of an hypothesis to a list of predefined distress calls. Each ASR hypothesis H_i is phonetized, every voice commands T_j is aligned to H_i using Levenshtein distance. The deletion, insertion and substitution costs were computed empirically while the cumulative distance $\gamma(i,j)$ between H_j and T_i is given by Equation 1.

$$\gamma(i,j) = d(T_i, H_j) + \\ min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (1)$$

The decision to select or not a detected sentence is then taken according a detection threshold on the aligned symbol score (phonems) of each identified call. This approach takes into account some recognition errors like word endings or light variations. Moreover, in a lot of cases, a miss-decoded word is phonetically close to the good one (due to the close pronunciation). From this the CER (Call Error Rate i.e., distress call error rate) is defined as:

$$\text{CER} = \frac{\text{Number of missed calls}}{\text{Number of calls}} \quad (2)$$

This measure was chosen because of the content of the corpus Cirdo-set used in this study. Indeed, this corpus is made of sentences and interjections. All sentences are calls for help, without any other kind of sentences like home automation orders or colloquial sentences, and therefore it is not possible to determine a false alarm rate in this framework.

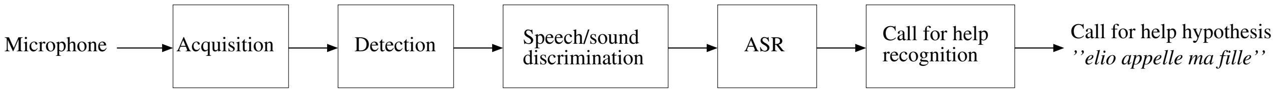

Microphone
Acquisition
Detection
Speech/sound discrimination
ASR
Call for help recognition
Call for help hypothesis
''elio appelle ma fille''

Figure 2: Architecture of the analysis system CIRDOX

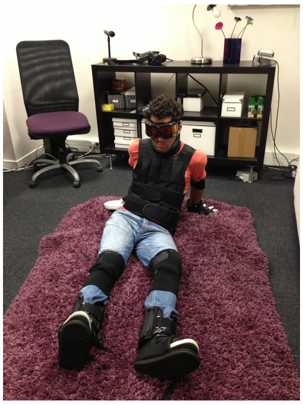


Figure 3: A young participant playing a fall scenario

3. Experimentation and results

3.1. Live Experiment

An experiment was run in the experimental platform of the LIG laboratory in a room whose setting corresponds to Figure 1 and equipped with a sofa, a carpet, 2 chairs, a table and e-lio. A Sennheiser SKM 300 G2 ME2 omnidirectional microphone was set on the cupboard. In these conditions, the microphone was at a distance of above 2 meters from the speaker (Distant speech conditions). The audio analysis system consisted in the CIRDOX software presented in Section 2 which was continuously recording and analysing the audio streams to detect the calls.

3.1.1. Scenarios and experimental protocol

The scenarios were elaborated after field studies made by the GRePS laboratory [17]. These studies allowed to specify the fall context, the movements during the fall as well as the person's reaction once on the floor. Phrases uttered during and after the fall were also identified "*Blast! What's happening to me? Oh shit, shit!*". The protocol was as follows [18]. Each participant was introduced to the context of the research and was invited to sign a consent form. The participants played four scenarios of fall, one blocked hip scenario and two other scenarios called "true-false" added to challenge the automatic detection of falls by the video analysis system. If the participant's age was under 60, he wore a simulator which hampered his mobility and reduced his vision and hearing to simulate aged physical conditions. Figure 3 shows a young participant wearing the simulator at the end of a fall scenario. The average experiment duration of an experiment was 2h 30min per person. This experiment was very tiring for the participants and it was necessary to include rehearsals before starting the recordings so that the participant felt comfortable and was able to fall securely.

3.1.2. Voice commands and distress calls

The sentences of the AD80 corpus [19] served as basis to develop the language model used by our system. This corpus was recorded by 43 elderly people and 52 non-aged pepole in our laboratory and in a nursing home to study the automatic recognition of speech uttered by aged speakers. This corpus is made of 81 casual sentences, 31 vocal commands for home automation and 58 distress sentences. An excerpt of these sentences in French is given Table 2, the distress sentences identified in the field study reported in section 3.1.1 were included in the corresponding part of AD80.

The utterance of some of these distress sentences were integrated into the scenarios with the exception of the two "true-false" scenarios.

3.1.3. Acquired data: Cirdo-set

In this paper we focus on the detection of the distress calls, therefore we don't consider the audio event detected and analyzed on the fly but only the full records of each scenario. These data sets were transcribed manually using transcriber [26] and the speech segments were then extracted for analysis.

The targeted participants were elderly people that were still able to play the fall scenarios securely. However, the recruitment of such kind of population was very difficult and a part of the participants was composed of people under 60 years old but they were invited to wear a special suit [18] which hampered their mobility and reduced their vision but without any effect on speech production. Overall, 17 participants were recruited (9 men and 8 women). Among them, 13 participants were under 60 and worn the simulator. The aged participants were between 61 and 83 years old.

When they played the scenarios, some participants produced sighs, grunts, coughs, cries, groans, pantings or throat clearings. These sounds were not considered during the annotation process. In the same way, speeches mixed with sound produced by the fall were ignored. At the end, each speaker uttered between 10 and 65 short sentences or interjections ("*ah*", "*oh*", "*aïe*", etc.) as shown Table 1.

Sentences were often close of those identified during the field studies ("*je peux pas me relever* - I can't get up", "*e-lio appelle du secours* - e-lio call for help", etc.), some were different ("*oh bein on est bien là tiens* - oh I am in a sticky situation"). In practice, participants cut some sentences (i.e., inserted a delay between "*e-lio*" and "*appelle ma fille* - call my daughter"), uttered some spontaneous sentences, interjections or non-verbal sounds (i.e., groan).

3.2. Off line experiments

The methods presented in Section 2 were run on the Cirdo-set corpus presented in Section 3.1.3.

The SGMM model presented in Section 2.2 was used as acoutic model. The *generic language model* (LM) was estimated from French newswire collected in the Gigaword corpus. It was 1- gram with 13,304 words. Moreover, to reduce the linguistic variability, a 3-gram domain language model, the *specialized language model* was learnt from the sentences used during the corpus collection described in Section 3.1.1, with 99 1-gram, 225 2-gram and 273 3-gram models. Finally, the lan-

Distress Sentence	*Home Automation Command*	*Casual Sentence*
Aïe aïe aïe ⋆	Appelle quelqu'un e-lio ⋆	Bonjour madame
Oh là ⋆	e-lio, appelle quelqu'un ⋆	Ça va très bien
Merde ⋆	e-lio tu peux appeler une ambulance	Où sont mes lunettes
Je suis tombé ⋆	e-lio tu peux téléphoner au SAMU	Le café est brûlant
Je peux pas me relever ⋆	e-lio, appelle du secours	J'ai ouvert la porte
Qu'est-ce qu'il m'arrive ⋆	e-lio appelle les secours	Je me suis endormi tout de suite
Aïe ! J'ai mal ⋆	e-lio appelle ma fille	Il fait soleil
Oh là ! Je saigne ! Je me suis blessé ⋆	e-lio appelle les secours	Ce livre est intéressant
Aidez-moi	e-lio appelle le SAMU !	Je dois prendre mon médicament
Au secours	e-lio appelle les pompiers !	J'allume la lumière

Table 2: Examples of sentences of the AD80 corpus (⋆ denotes a sentence identified during the sociological study)

Spk.	Age	Sex	Nb. of interjections or short sentences	
			All	*Distress*
S01	30	M	22	14
S02	-	-	-	
S03	24	F	16	15
S04	83	F	65	53
S05	29	M	24	21
S06	64	F	23	19
S07	61	M	23	21
S08	44	M	25	15
S09	16	M	32	21
S10	16	M	19	15
S11	52	M	12	12
S12	28	M	15	12
S13	66	M	24	21
S14	52	F	23	21
S15	23	M	20	19
S16	40	F	29	27
S17	40	F	24	21
S18	25	F	17	14
Total	40.76		413	341

Table 1: Composition of the audio corpus Cirdo-set

guage model was a 3-gram-type which resulted from the combination of the *generic LM* (with a 10% weight) and the *specialized LM* (with 90% weight). This combination has been shown as leading to the best WER for domain specific application [27]. The interest of such combination is to bias the recognition towards the domain LM but when the speaker deviates from the domain, the general LM makes it possible to avoid the recognition of sentences leading to "false-positive" detection.

Results on manually annotated data are given Table 3. The most important performance measures are the Word Error Rate (WER) of the overall decoded speech and those of the specific distress calls as well as the Call Error Rate (CER: c.f. equation 2). Considering distress calls only, the average WER is 34.0% whereas it a 39.3% when all interjections and sentences are taken into account.

Unfortunately and as mentionned above, the used corpus doesn't allow the détermine a False Alarm Rate. Previous studies based on the AD80 corpus showed recall, precision and F-measure equal to 88.4%, 86.9% and 87.2% [19]. Nevertheless, this corpus was recorded in very different conditions, text reading in a studio, in contrary of those of Cirdo-set.

Spk.	WER (%)		CER	Spk.	WER (%)		CER
	All	*Distress*	(%)		*All*	*Distress*	(%)
S01	45.0	39.1	27.8	S11	21.3	17.0	16.7
S03	41.4	44.4	40.0	S12	30.8	25.0	25.0
S04	51.9	49.6	34.0	S13	45.9	43.6	23.8
S05	19.1	15.4	14.3	S14	67.0	54.8	50.0
S06	39.2	34.3	26.3	S15	21.5	19.5	5.3
S07	21.2	20.3	28.6	S16	14.9	11.76	7.4
S08	61.8	50.8	20.0	S17	21.4	22.4	19.0
S09	49.4	41.2	33.3	S18	57.7	44.9	71.4
S10	24.5	22.4	14.3	All	39.3	34.0	26.8

Table 3: Word and Call Error Rate for each participant

On average, CER is equal to 26.8% with an important disparity between the speakers.

4. Discussion

These results are quite different from those obtained with the AD80 corpus (with aged speakers and speaker adaptation): WER was 14.5% [19]. There are important differences between the recording conditions used for AD80 and for the Cirdo-set corpus used in our study that can explain this performance gap:

- AD80 is made of readings by speakers sitting in comfortable position in front of a PC and the microphone ;
- AD80 was recorded in nearest conditions in comparison with distant setting for Cirdo-set ;
- Cirdo-set was recorded by participants who fell on the floor or that are blocked on the sofa. They were encouraged to speak in the same way that they would speak if they would be really put in these situations. Obviously, we obtained expressive speech, but there is no evidence that the pronunciation would be the same as in real conditions of a fall or a blocked hip.

Regarding the CER, its global value 26.8% shows that 74.2% of the calls were correctly recognized ; furthermore, at the exception of one speaker (CER=71.4%), CER is always below 50% consequently more than 50% of the calls were recognized. For 6 speakers, CER was below 20%. This suggests that a distress call could be detected if the speaker is able to repeat his call two or three times. However, if the system did not identify the first distress call because the person's voice is altered by the stress, it is likely that this person will fill more and more stress and as a consequence future calls would be more difficult to identify. In a same way, our corpus was recorded in realistic conditions but not in real conditions and frail elderly people may not be adequately simulated by healthy human adults. A relatively small number of missed distress calls could render the system unacceptable for use amongst the potential user and therefore some efforts in this regard would need to be pursued.

5. Conclusion and perspectives

This study is focused on the framework of automatic speech recognition applications in smart homes, that is in distant speech conditions and especially in realistic conditions very different from those of corpus recording when the speaker is reading a text.

Indeed in this paper, we presented the Cirdo-set corpus made of distress calls recorded in distant speech conditions and in realistic conditions in case of fall or blocked hip. The WER obtained at the output of the dedicated ASR was 36.3% for the distress calls. Thanks to a filtering of the ASR hypothesis at phonetic level, more than 70% of the calls were detected.

These results obtained in realistic conditions gives a fairly accurate idea of the performances that can be achieved with state of the art ASR systems for end user and specific applications. They were obtained in the particular case of the recognition of distress calls but they can be extended to other applications in which expressive speech may be considered because it is inherently present.

As stated above, obtained results are not sufficient to allow the system use in real conditions and two research ideas can be considered. Firstly, speech recognition performances may be improved thanks to acoustic models adapted to expressive speech. This may be achieved to the record of corpora in real conditions but this is a very difficult task. Secondly, it may be possible to recognize the repetition, at regular intervals, of speech events that are phonetically similar. This last method does not request the good recognition of the speech. Our future studies will address this problem.

6. Acknowledgements

This work is part of the CIRDOproject founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-10-TECS-012). The authors would like to thank the persons who agreed to participate in the recordings.

7. References

[1] K. K. B. Peetoom, M. A. S. Lexis, M. Joore, C. D. Dirksen, and L. P. De Witte, “Literature review on monitoring technologies and their outcomes in independently living elderly people,” *Disability and Rehabilitation: Assistive Technology*, pp. 1–24, 2014.

[2] L. C. D. Silva, C. Morikowa, and I. M. Petra, “State of the art of smart homes,” *Engineering Applications of Artificial Intelligence*, no. 25, pp. 1313–1321, 2012.

[3] M. Chan, D. Estève, C. Escriba, and E. Campo, “A review of smart homes- present state and future challenges,” *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 55–81, 2008.

[4] L. De Silva, C. Morikawa, and I. Petra, “State of the art of smart homes,” *Engineering Applications of Artificial Intelligence*, vol. 25, no. 7, pp. 1313–1321, 2012.

[5] Q. Ni, A. B. García Hernando, and I. P. de la Cruz, “The elderly’s independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development,” *Sensors*, vol. 15, no. 5, pp. 11 312–11 362, 2015.

[6] F. Bloch, V. Gautier, N. Noury, J. Lundy, J. Poujaud, Y. Claessens, and A. Rigaud, “Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects,” *Annals of Physical and Rehabilitation Medicine*, vol. 54, pp. 391–398, 2011.

[7] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, “Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects,” *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.

[8] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, “homeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition,” in *4th Workshop on Speech and Language Processing for Assistive Technologies*, 2014.

[9] M. Vacher, F. Portet, A. Fleury, and N. Noury, “Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges,” *International Journal of E-Health and Medical Communications*, vol. 2, no. 1, pp. 35–54, 2011.

[10] W. Li, J. Glass, N. Roy, and S. Teller, “Probabilistic dialogue modeling for speech-enabled assistive technology,” in *SLPAT 2013*, 2013, pp. 67–72.

[11] J. F. Gemmeke, B. Ons, N. Tessema, H. Van Hamme, J. Van De Loo, G. De Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. Van Den Broeck, P. Karsmakers, and B. Vanrumste, “Self-taught assistive vocal interfaces: an overview of the ALADIN project,” in *Interspeech 2013*, 2013, pp. 2039–2043.

[12] P. Milhorat, D. Istrate, J. Boudy, and G. Chollet, “Hands-free speech-sound interactions at home,” in *EUSIPCO 2012*, Aug. 2012, pp. 1678 –1682.

[13] M. Matassoni, R. F. Astudillo, A. Katsamanis, and M. Ravanelli, “The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones,” in *Interspeech 2014*, Sep. 2014, pp. 1613–1617.

[14] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux, and P. Chahuara, “Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation,” *ACM Transactions on Accessible Computing, Special Issue on Speech and Language Processing for AT (Part 3)*, vol. 7, no. 2, (in press), 36 pages.

[15] M. Vacher, B. Lecouteux, D. Istrate, T. Joubert, F. Portet, M. Sehili, and P. Chahuara, “Evaluation of a Real-Time Voice Order Recognition System from Multiple Audio Channels in a Home,” in *Interspeech 2013*, Aug. 2013, pp. 2062–2064.

[16] M. Vacher, B. Lecouteux, and F. Portet, “Multichannel Automatic Recognition of Voice Command in a Multi-Room Smart Home : an Experiment involving Seniors and Users with Visual Impairment,” in *Interspeech 2014*, Sep. 2014, pp. 1008–1012.

[17] M. Bobillier Chaumon, F. Cros, B. Cuvillier, C. Hem, and E. Codreanu, “Concevoir une technologie pervasive pour le maintien à

domicile des personnes âgées : la détection de chutes dans les activités quotidiennes," in *Activités Humaines, Technologies et bien-être, Congrès EPIQUE (Psychologie Ergonomique)*, Belgique - Bruxelles, July 2013, pp. 189–199.

[18] S. Bouakaz, M. Vacher, M.-E. Bobillier-Chaumon, F. Aman, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, E. Desserée, P. Traineau, J.-P. Vimon, and T. Chevalier, "CIRDO: Smart companion for helping elderly to live at home for longer," *Innovation and Research in BioMedical engineering (IRBM)*, vol. 35, no. 2, pp. 101–108, Mar. 2014.

[19] F. Aman, M. Vacher, S. Rossato, and F. Portet, "Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences," in *The 7th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2013*, Cluj-Napoca, Romania, Oct. 2013, pp. 177–184.

[20] M. Vacher, D. Istrate, and J. Serignat, "Sound detection and classification through transient models using wavelet coefficient trees," in *Proc. 12th European Signal Processing Conference*, S. LTD, Ed., Vienna, Austria, sep. 2004, pp. 1171–1174.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[22] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The Sweet-Home speech and multimodal corpus for home automation interaction," in *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014, pp. 4499–4506.

[23] F. Aman, "Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile," Ph.D. dissertation, Université de Grenoble, Ecole doctorale MSTII, 2014.

[24] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model—a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404 – 439, 2011.

[25] L. Zouari and G. Chollet, "Efficient gaussian mixture for speech recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4, 2006, pp. 294–297.

[26] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1-2, pp. 5–22, 2001.

[27] B. Lecouteux, M. Vacher, and F. Portet, "Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions," in *Proc. InterSpeech*, 2011, pp. 2273–2276.

A Comparison of Manual and Automatic Voice Repair for Individual with Vocal Disabilities

Christophe Veaux, Junichi Yamagishi, Simon King

Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
{cveaux, jyamagis}@inf.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

When individuals lose the ability to produce their own speech, due to degenerative diseases such as motor neurone disease (MND) or Parkinson's, they lose not only a functional means of communication but also a display of their individual and group identity. In order to build personalized synthetic voices, attempts have been made to capture the voice before it is lost, using a process known as voice banking. But, for some patients, the speech deterioration frequently coincides or quickly follows diagnosis. Using HMM-based speech synthesis, it is now possible to build personalized synthetic voices with minimal data recordings and even disordered speech. The power of this approach is that it is possible to use the patient's recordings to adapt existing voice models pre-trained on many speakers. When the speech has begun to deteriorate, the adapted voice model can be further modified in order to compensate for the disordered characteristics found in the patient's speech, we call this process "voice repair". In this paper we compare two methods of voice repair. The first method follows a trial and error approach and requires the expertise of a speech therapist. The second method is entirely automatic and based on some a priori statistical knowledge. A subjective evaluation shows that the automatic method achieves similar results than the manually controlled method.

Index Terms: HTS, Speech Synthesis, Voice Banking, Voice Reconstruction, Voice Output Communication Aids, MND.

1. Introduction

Degenerative speech disorders have a variety of causes that include Multiple Sclerosis, Parkinson's, and Motor Neurone Disease (MND) also known in the USA as Amyotrophic Lateral Sclerosis (ALS). MND primarily affects the motor neurones in the brain and spinal cord. This causes a worsening muscle weakness that leads to a loss of mobility and difficulties with swallowing, breathing and speech production. Initial symptoms may be limited to a reduction in speaking rate, an increase of the voice's hoarseness, or an imprecise articulation. However, at some point in the disease progression, 80 to 95% of patients are unable to meet their daily communication needs using their speech [1]. As speech becomes difficult to understand, these individuals may use a voice output communication aid (VOCA). These devices consist of a text entry interface such as a keyboard, a touch screen or an eye-tracker, and a text-to-speech synthesizer that generates the corresponding speech. However, when individuals lose the ability to produce their own speech, they lose not only a functional means of communication but also a display of their individual and social identity through their vocal characteristics.

Current VOCAs are not ideal as they are often restricted to a limited set of impersonal voices that are not matched to the age or accent of each individual. Feedback from patients, careers and patient societies has indicated that there is a great unmet need for personalized VOCAs as the provision of personalized voice is associated with greater dignity and improved self-identity for the individual and their family [2]. In order to build personalized VOCAs, several attempts have been made to capture the voice before it is lost, using a process known as voice banking. One example of this approach is ModelTalker [3], a free voice building service that can be used from any home computer in order to build a synthetic voice based on diphone concatenation, a technology developed in the 1980s. The user of this service has to record around 1800 utterances in order to fully cover the set of diphones and the naturalness of the synthetic speech is rather low. Cereproc [4] has provided a voice building service for individuals, at a relatively high cost, which uses unit selection synthesis, and is able to generate synthetic speech of increased naturalness. However, these speech synthesis techniques require a large amount of recorded speech in order to build a good quality voice. Moreover the recorded speech data must be as intelligible as possible, since the data recorded is used directly as the voice output. This requirement makes such techniques more problematic for those patients whose voices have started to deteriorate. Therefore, there is a strong motivation to improve the voice banking and voice building techniques, so that patients can use their own synthetic voices, even if their speech is already disordered at the time of recordings. A first approach is to try to separate out the disorders from the recorded speech. In this way, Rudzicz [5] has proposed a combination of several speech processing techniques. However, some disorders cannot be simply filtered out by signal processing techniques and a model-based approach seems more appropriate. Kain [6] has proposed a voice conversion framework for the restoration of disordered speech. In its approach, the low-frequency spectrum of the voiced speech segment is modified according to a mapping defined by a Gaussian mixture model (GMM) learned in advance from a parallel dataset of disordered and target speech. The modified voiced segments are then concatenated with the original unvoiced speech segments to reconstruct the speech. This approach can be seen as a first attempt of model-based voice reconstruction although it relies only on a partial modeling of the voice components. A voice building process using the hidden Markov model (HMM)-based speech synthesis technique has been investigated to create personalized VOCAs [7-10]. This approach has been shown to produce high quality output and offers two major advantages over existing methods for voice banking and voice building. First, it is possible to use existing speaker-independent voice models pre-trained over a number of speakers and to adapt them towards a target speaker. This process known as speaker adaptation [11] requires only a very

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 130–133,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

small amount of speech data. The second advantage of this approach is that we can control and modify various components of the adapted voice model in order to compensate for the disorders found in the patient's speech. We call this process "voice repair". In this paper, we compare different strategies of voice repair using the HMM-based synthesis framework. The first method follows a trial and error approach and requires the expertise of a speech therapist. The second method is entirely automatic and based on some a priori statistical knowledge.

2. HMM-Based Speech Synthesis

Our voice building process is based on the state-of-the-art HMM-based speech synthesizer, known as HTS [12]. As opposed to diphone or unit-selection synthesis, the HMM-based speech synthesizer does not use the recorded speech data directly as the voice output. Instead it is based on a vocoder model of the speech and the acoustic parameters required to drive this vocoder are represented by a set of statistical models. The vocoder used in HTS is STRAIGHT and the statistical models are context-dependent hidden semi-Markov models (HSMMs), which are HMMs with explicit state duration distributions. The state output distributions of the HSMMs represent three separate streams of acoustic parameters that correspond respectively to the fundamental frequency (logF0), the band aperiodicities and the mel-cepstrum, including their dynamics. For each stream, additional information is added to further describe the temporal trajectories of the acoustic parameters, such as their global variances over the learning data. Finally, separate decision trees are used to cluster the state durations probabilities and the state output probabilities using symbolic context information at the phoneme, syllable, word, and utterance level. In order to synthesize a sentence, a linguistic analyser is used to convert the sequence of words into a sequence of symbolic contexts and the trained HSMMs are invoked for each context. A parameter-generation algorithm is then used to estimate the most likely trajectory of each acoustic parameter given the sequence of models. Finally the speech is generated by the STRAIGHT vocoder driven by the estimated acoustic parameters.

3. Speaker Adaptation

One advantage of the HMM-based speech synthesis for voice building is that the statistical models can be estimated from a very limited amount of speech data thanks to speaker adaptation. This method [9] starts with a speaker-independent model, or "**average voice model**", learned over multiple speakers and uses model adaptation techniques drawn from speech recognition such as maximum likelihood linear regression (MLLR), to adapt the speaker independent model to a new speaker. It has been shown that using 100 sentences or approximately 6-7 minutes of speech data is sufficient to generate a speaker-adapted voice that sounds similar to the target speech [7]. In the following of this paper we refer the speaker-adapted voices as "**voice clones**". This provides a much more practical way to build a personalized voices for patients. For instance, it is now possible to construct a synthetic voice for a patient prior to a laryngectomy operation, by quickly recording samples of their speech [8]. A similar approach can also be used for patients with neurodegenerative diseases such as MND. However, we do not want to reproduce the symptoms of a vocal problem if the speech has already been disordered at the time of the recording. This is the aim of the voice repair methods introduced in the section 5 of this paper.

4. Database of Voice Donors

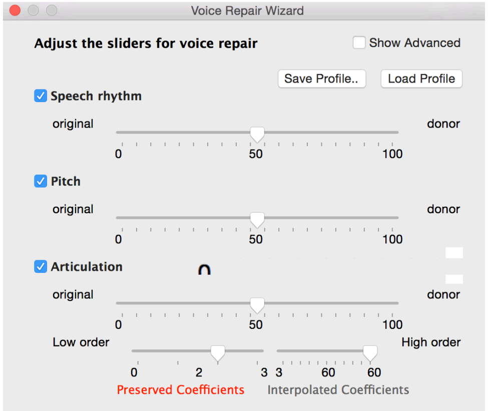
Ideally, the average voice model used for the speaker adaptation should be close to the vocal identity of the patient. On the other hand, a minimum number of speakers are necessary to train robust average voice models. Therefore, we have created a database of more than 900 healthy voice donors with various accents (Scottish, Irish, Other UK). Each speaker recorded about one hour of speech (400 sentences). This database of healthy voices is first used to create the average voice models used for speaker adaptation. Ideally, the average voice model should be close to the vocal identity of the patient and it has been shown that gender and regional accent are the most influent factors in speaker similarity perception [13]. Therefore, the speakers are clustered according to their gender and their regional accent in order to train specific average voice models. A minimum of 10 speakers is required in order to get robust average voice models. Furthermore, the database is also used to select a reference donor for the voice repair procedures described in section 5. The voice repair is most successful when the reference donor is as close as possible to the patient in terms of vocal identity.

5. Voice Repair

Some individuals with neurodegenerative disease may already have speech symptoms at the time of the recording. In that case, the speaker adaptation process will also replicate these symptoms in the speaker-adapted voice. Therefore we need to remove speech disorders from the synthetic voice, so that it sounds more natural and more intelligible. Repairing synthetic voices is conceptually similar to the restoration of disordered speech mentioned in Section 1, but we can now exploit the acoustic models learned during the training and the adaptation processes in order to control and modify various speech features. This is the second major advantage of using HMM-based speech synthesis. In particular, HTS has statistically independent models for duration, log-F0, band aperiodicity and mel-cepstrum. This allows the substitution of some models in the patient's speaker-adapted voice by that of a well-matched healthy voice or an average of multiple healthy voices. For example, patients with MND often have a disordered speaking rate, contributing to a loss of the speech intelligibility. The substitution of the state duration models enables the timing disruptions to be regulated at the phoneme, word, and utterance levels. Furthermore, MND speakers often have breathy or hoarse speech, in which excessive breath through the glottis produces unwanted turbulent noise. In such cases, we can substitute the band aperiodicity models to produce a less breathy or hoarse output. In the following part of this section, we present two different methods of model substitution. The first one is manually controlled whereas the second one is automatic.

5.1. Manual voice repair

In the manual approach, a speech therapist first selects a reference voice among all the available voices with same accent, gender and age range than the patient. Then the models of this reference voice are used to correct some of the patient's voice models. This correction is based on mean and variance interpolation between models. A graphical interface allows the speech therapist to control the amount of interpolation between the patient's voice models and the reference voice models as illustrated in Figure 1.



Voice Repair Wizard

Adjust the sliders for voice repair □ Show Advanced

Save Profile.. | Load Profile

☑ **Speech rhythm**

original — 0 — 50 — 100 — donor

☑ **Pitch**

original — 0 — 50 — 100 — donor

☑ **Articulation**

original — 0 — 50 — 100 — donor

Low order — 0 — 2 — 3 | 3 — 60 — 60 — High order

Preserved Coefficients | Interpolated Coefficients

Figure 1: Graphical interface for model interpolation.

The following models and information can be interpolated:

- Duration
- Dynamics coefficients of the log-F0
- Dynamics coefficients of the mel-cepstrum
- Low-order coefficients of the mel-cepstrum
- High-order coefficients of the mel-cepstrum

The voiced/unvoiced weights and aperiodicity models are simply substituted since their impact on voice identity is rather limited and their replacement of will fix the breathiness disorders. The interpolation of the high order static coefficients and the dynamics coefficients of the mel-cepstrum will help to reduce the articulation disorders without altering the timbre. The interpolation of the dynamics coefficients of the log-F0 will help to regulate the prosodic disorders such as monotonic F0. Finally the global variances of all the parameters are also simply substituted. We will refer to this method as the **manual repair**.

5.2. Automatic voice repair

The manual voice repair requires a lot of expertise from the speech therapist, as it is a trial and error approach. Therefore, we aim to replace it by a fully automated voice repair procedure. We measure the Kullback-Leibler distance (KLD) between the models of the patient voice and the models of the reference voice as illustrated in Figure 2. Then the likelihood of each of the measured distance is evaluated given the statistical distribution of KLD distances between healthy voice models of similar accent, gender and age band. The likelihood values are used to control the interpolation between the patient and reference voice models. For instance, if the likelihood of the KLD distance for a given model of the patient voice is very low, the corresponding model of the reference voice is used to replace it in the patient voice. The reference voice model is also selected automatically as the one that maximizes the likelihood of the patient recording data.

6. Experiment

The manual and automatic voice repair methods presented in Section 5 were evaluated for the case of a MND patient. This patient was a 45 years old Scottish male that we recorded twice. A first recording of one hour (500 sentences) has been made just after diagnosis when he was at the very onset of the disease.

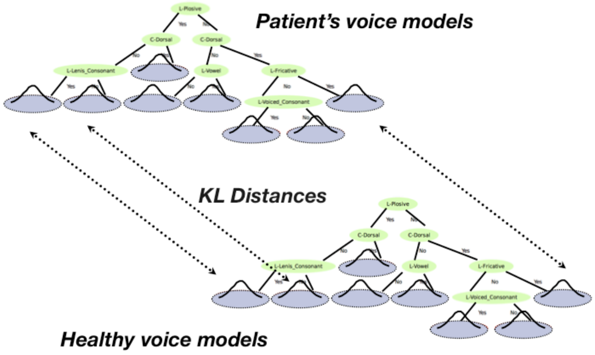


Figure 2: Graphical interface for model interpolation.

At that time, his voice did not show any disorders and could still be considered as "healthy". A second recording of 15 minutes (50 sentences) has been made 10 months later. He has then acquired some speech disorders typically associated with MND, such as excessive hoarseness and breathiness, disruption of speech fluency, reduced articulation and monotonic prosody. These two recordings were used separately as adaptation data in order to create two speaker-adapted voices from the same male-Scottish average voice model. The synthetic voice created from the first recording of the patient ("healthy" speech) was used as the reference voice for the subjective evaluations. This choice of a synthetic voice as reference instead of the natural recordings was done to avoid any bias due to the loss of quality inherent to the synthesis. Two different reconstructed voices were created from the second recording of the patient ("impaired" speech) using the manual and the automatic voice repair methods respectively. In order to evaluate the voice repair methods, two subjective tests were conducted. The first one assesses the intelligibility of the reconstructed voices whereas the second one measures their similarity with synthetic voice created from "healthy" speech of the patient. We also included the synthetic voices of the donors selected for the manual and the automatic voice repair in the similarity test. All the synthetic voices used in the experiment are summarized in Table 1.

Voice	*Description*
MD	Voice of donor used in manual voice repair
AD	Voice of donor used in automatic voice repair
HC	Voice clone of the "**healthy**" speech (1st recording)
IC	Voice clone of the "**impaired**" speech (2nd recording)
IR_v1	Reconstructed voice using **manual voice repair**
IR_v2	Reconstructed voice using **automatic voice repair**

Table 1: Voices compared in the evaluation tests.

6.1. Listening Intelligibility Test

The same 50 semantically unpredictable sentences were synthesized for each of the voices created from the patient's recordings (see Table 1). The resulting 200 synthesized samples were divided into 4 groups such that each voice is represented by 10 samples in a group. A total of 40 native English participants were asked to transcribe the synthesized samples, with 10 participants for each group. Within each group, the samples were presented in random order for each participant. The participants performed the test with headphones. The transcriptions were evaluated by measuring the word error rate (WER).

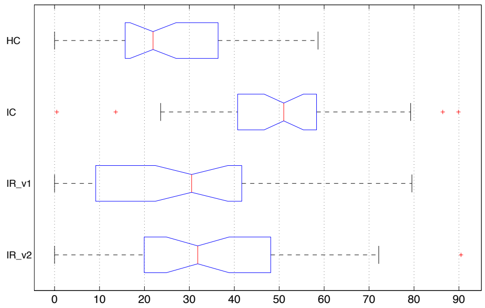


Figure 3: *Word Error Rate (mean and standard deviation)*

6.2. Speaker Similarity Test

The same test sentence “People look, but no one ever finds it.” was synthesized for each of the voices in Table 1. Participants were asked to listen alternatively to the reference voice (HC) and to the same sentence synthesized with one of the other voices. The presentation order of the voice samples was randomized. The participants have been asked to rate the similarity in terms of speaker identity between the tested voice and the reference (HC) on a 5-point scale (1: Very dissimilar, 2: Dissimilar, 3: Quite Similar, 4: Very similar; and 5: Identical). A total of 40 native English speakers performed the test using headphones.

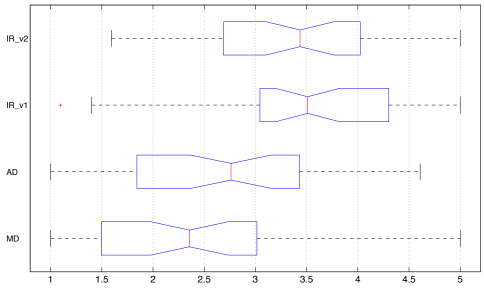


Figure 4: *Similarity to the reference voice HC on a MOS-scale (mean and standard deviation)*

7. Results and Discussion

The resulting average WERs for the intelligibility test are shown in Figure 2. We are not interested here in the absolute values of the WER but in their relative values compared to the healthy voice HC. As expected, the synthetic voice IC created from the “impaired” speech has a high WER. Both manual and automatic voice repair succeeds in removing some articulation disorders from the synthetic speech as we can see a significant decrease of WER. The manual voice repair yields to slightly lower WER than the automatic voice repair although the difference is not significant. The results of the similarity test are shown in Figure 3. The first important result is that the reconstructed voices are still considered more similar to the patient’s voice than the closest voice donors (MD and AD) used for the voice repair. This means that both voice repair methods manage to preserve the voice identity to a certain extent. The manual voice repair is performing slightly better than the automatic method but the difference is not significant (p-value ~ 1.e-2).

8. Conclusions

HMM-based speech synthesis has two clear advantages for the creation of personalized voices for people with disordered speech: speaker adaptation and improved control. Speaker adaptation allows the creation of a voice clone with a limited amount of data. Then the structure of the acoustic models can be modified to repair the synthetic speech. We have presented here two different strategies for voice reconstruction. The first one is manual and requires the expertise of a speech therapist whereas the second one is fully automated. The evaluation of these methods demonstrates that: a) it is possible to improve the intelligibility of a disordered synthetic speech while retaining its vocal identity; b) the automatic voice repair performs almost as well as the manual voice repair. The reconstruction strategies presented here have been designed for MND patients, but their principle could be easily generalized to any other degenerative or acquired speech disorder.

9. References

[1] Doyle, M. and Phillips, B. (2001), “Trends in augmentative and alternative communication use by individuals with amyotrophic lateral sclerosis,” *Augmentative and Alternative Communication* 17 (3): pp.167–178.

[2] Murphy, J. (2004), “I prefer this close’: Perceptions of AAC by people with motor neurone disease and their communication partners. *Augmentative and Alternative Communication*, 20, 259-271.

[3] Yarrington, D., Pennington, C., Gray, J., & Bunnell, H. T. (2005), “A system for creating personalized synthetic voices,” *Proc. of ASSETS.*

[4] http://www.cereproc.com/

[5] Rudzicz, F. (2011) “Production knowledge in the recognition of dysarthric speech”, PhD thesis, University of Toronto.

[6] Kain, A.B., Hosom, J.P. Niu X., van Santen J.P.H., Fried-Oken, M., and Staehely, J., (2007) “Improving the intelligibility of dysarthric speech,” Speech Communication, 49(9), pp743–759.

[7] Creer, S., Green, P., Cunningham, S., & Yamagishi, J. (2010) “Building personalized synthesized voices for individuals with dysarthia using the HTS toolkit,” IGI Global Press, Jan. 2010.

[8] Khan, Z. A., Green P., Creer, S., & Cunningham, S. (2011) “Reconstructing the Voice of an Individual Following Laryngectomy,” Augmentative and Alternative Communication.

[9] Veaux, C., Yamagishi, J., King, S. (2011) “Voice Banking and Voice Reconstruction for MND patients,” *Proceedings of ASSETS.*

[10] Veaux, C., Yamagishi, J., King, S. (2012) “Using HMM-based Speech Synthesis to Reconstruct the Voice of Individuals with Degenerative Speech Disorders,” *Interspeech*, Portland, USA.

[11] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. & Isogai, J. 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. on ASL, 17, 66-83.

[12] Zen, H., Tokuda, K., & Black, A. (2009) “Statistical parametric speech synthesis, Speech Communication,” 51, pp.1039-1064.

[13] Dall, R., Veaux, C., Yamagishi, J. & King, S. (2012) “Analysis of speaker clustering strategies for HMM-based speech synthesis,” *Proc. Interspeech*, Portland, USA.

Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias

Maria Yancheva[1], *Kathleen Fraser*[1], *Frank Rudzicz*[1,2]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Toronto Rehabilitation Institute-UHN, Toronto, Canada

yancheva@cs.toronto.edu, kfraser@cs.toronto.edu, frank@cs.toronto.edu

Abstract

We use a set of 477 lexicosyntactic, acoustic, and semantic features extracted from 393 speech samples in DementiaBank to predict clinical MMSE scores, an indicator of the severity of cognitive decline associated with dementia. We use a bivariate dynamic Bayes net to represent the longitudinal progression of observed linguistic features and MMSE scores over time, and obtain a mean absolute error (MAE) of 3.83 in predicting MMSE, comparable to within-subject interrater standard deviation of 3.9 to 4.8 [1]. When focusing on individuals with more longitudinal samples, we improve MAE to 2.91, which suggests at the importance of longitudinal data collection.

Index Terms- Alzheimer's disease, dementia, Mini-Mental State Examination (MMSE), dynamic Bayes network, feature selection

1. Introduction

Research into the early assessment, pathogenesis, and progression of dementia is becoming increasingly important, as the proportion of people it affects grows every year. Alzheimer's disease (AD), the most common type of dementia, affects more than half of the population above 80 years of age and its impact on society is expected to grow as the "baby boomer" generation ages [2, 3, 4].

There is no single laboratory test that can identify dementia with absolute certainty. Typically, probable dementia is diagnosed using the Mini Mental State Examination (MMSE), which provides a score on a scale of 0 (greatest cognitive decline) to 30 (no cognitive decline), based on a series of questions in five areas: orientation, registration, attention, memory, and language [5]. While MMSE provides a unified scale for measuring the severity of the disease, it can be time-consuming and relatively costly, often requiring a trained neuropsychologist or physician to administer the test in a clinical setting.

Changes in cognitive ability due to neurodegeneration associated with AD lead to a progressive decline in memory and language quality. Patients experience deterioration in sensory, working, declarative, and non-declarative memory, which leads to a decrease in the grammatical complexity and lexical content of their speech [6]. Such changes differ from the pattern of decline expected in older adults [6], which suggests that temporal changes in linguistic features can aid in disambiguation of healthy older adults from those with dementia.

Some previous work used machine learning classifiers with linguistic features for two-class separation of patients with AD from controls (see section 1.1), but there appears to be no previous research that has used them to infer a clinical score for dementia — an indicator of the degree of cognitive decline. The present work uses a set of automatically-extracted lexicosyntactic, acoustic, and semantic (LSAS) features for estimating continuous MMSE scores on a scale of 0 to 30, using a dynamic Bayes network for representing relationships between observed linguistic measures and underlying clinical scores.

Since dynamic changes in linguistic ability in patients with AD differ from those in typical healthy older adults [6], we hypothesize that considering speech samples over time would aid in estimating underlying cognitive status. Previous studies analyzing dynamic progression of language features in patients with AD did not employ machine learning techniques, and are characterized by a small number of subjects (between 3 and 6) and a limited set of features that do not include acoustics. The present work improves on these analyses by extracting LSAS features from a relatively large collection of longitudinal speech, in order to estimate MMSE scores.

1.1. Related Work

Previous work has explored the use of lexicosyntactic features for identifying individuals with AD from controls. Orimaye *et al.* [7] used DementiaBank[1], one of the largest existing datasets of pathological speech [8], to perform binary classification of 242 patients with dementia and 242 controls; a support vector machine classifier achieved their best F-measure of 0.74 [7]. Another experiment by Jarrold *et al.* collected spontaneous speech data from 9 controls, 9 patients with AD, and 30 patients with frontotemporal lobar degeneration (FTLD) [9]. A multi-layer perceptron model obtained classification accuracy of 88% on a two-class task (AD:controls, and FTLD:controls), and 80% on a three-class task (AD:FTLD:controls).

While these studies have obtained promising results in classifying patients with dementia based on linguistic features, there is limited work modelling the progression of such features over time. Le *et al.* [10] examined the longitudinal changes in a small set of hand-selected lexicosyntactic measures, such as vocabulary size, repetition, word class deficit, and syntactic complexity, in 57 novels of three British authors written over a period of several decades. They found statistically significant lexical deterioration in Agatha Christie's work evidenced by vocabulary impoverishment and a pronounced increase in word repetitions [10], but the measures for syntactic complexity did not yield conclusive results. A similar analysis performed by Sundermann examined the progression of a small set of lexicosyntactic features, such as length, frequency, and vocabulary measures in 6 patients with AD or mild cognitive impairment (MCI), with a minimum of 3 longitudinal samples in DementiaBank [11]. Analysis of the features over time did not reveal con-

[1] http://talkbank.org/DementiaBank/

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 134–139,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

clusive patterns; Sundermann suggested that the limited sample size and feature set selection may be the cause. Neither study involved acoustics or machine learning techniques.

2. Methodology

2.1. Data

We use data from DementiaBank, a large dataset of speech produced by people with dementia (including probable AD, possible AD, vascular dementia, and MCI) and healthy older adults, recorded longitudinally at the University of Pittsburgh's Alzheimer's Disease Research Center [8]. Annual visits with each subject consist of a recording of speech data, its textual transcription, and an MMSE score. Subjects have a variable number of longitudinal samples ($min = 1$, $max = 5$, $M = 1.54$, $SD = 0.79$). Each speech sample consists of a verbal description of the Boston Cookie Theft picture, which typical lasts about a minute. We partition subjects between controls (CT) and those with probable AD, possible AD or MCI (or, collectively,"AD" [2]). Considering only subjects with associated MMSE scores, the working set consists of 393 speech samples from 255 subjects (165 AD, 90 CT).

2.2. Features

Three major types of features are extracted from the speech samples and their transcriptions: (1) *lexicosyntactic measures*, extracted from syntactic parse trees constructed with the Brown parser and POS-tagged transcriptions of the narratives [12, 13, 14, 15, 16]; (2) *acoustic measures*, including the standard Mel-frequency cepstral coefficients (MFCCs), formant features, and measures of disruptions in vocal fold vibration regularity [17]; and (3) *semantic measures*, pertaining to the ability to describe concepts and objects in the Cookie Theft picture. The full list of features, along with their major type and subtype, is shown in Table 1.

2.3. Feature Analysis

Two feature selection methods are used to identify the most informative features for disambiguating AD from CT. Since the MMSE score is a measure of the progression of cognitive impairment and is used to distinguish AD from CT generally, we hypothesize that highly discriminating features of the two groups would also be good predictors of MMSE. This is quantified by Spearman's rank-order correlation between the most informative features and the MMSE score, ρ_{MMSE}, shown in Table 2.

The first feature ranking method is a two-sample t-test ($\alpha = 0.001$, two-tailed) which quantifies the significance of the difference in each feature value between the two classes; the features are ordered by increasing p-value. Table 2 shows the type and p-value of the top 10 features, along with their correlation with MMSE. Control subjects use longer utterances, more gerund + prepositional phrase constructions (VP→ VBG PP, e.g., *standing on the chair*), more content words such as noun phrases (NP) and verbs, and are more likely to talk about what they see through the window (info_window), which is in the background of the scene (e.g., *it seems to be summer out*). On the other hand, subjects with AD use more words not found in the dictionary (NID), and more function words such as pronouns (PRP). Honoré's statistic measures lexical richness, extending type-token ratio, which is decreased in AD. These findings are consistent with expectations.

Table 2: The top 10 features selected by a two-sample t-test ($\alpha = 0.001$, two-tailed) as the most informative discriminators of AD versus CT. ρ_{MMSE} is Spearman's rank-order correlation coefficient between the given feature and the MMSE score. The features in **bold** are among the top 10 selected by mRMR.

Feature	Feature type	p	ρ_{MMSE}
avelength	lexicosyntactic	1.24E-13	0.3837
VP → VBG PP	lexicosyntactic	1.90E-13	0.3757
NID	lexicosyntactic	3.23E-11	-0.3712
NP → DT NN	lexicosyntactic	1.12E-10	0.3438
NP → PRP	lexicosyntactic	2.14E-10	-0.3186
prp_ratio	lexicosyntactic	1.16E-09	-0.3089
honoré	lexicosyntactic	2.53E-09	0.3400
verbs	lexicosyntactic	4.81E-09	0.2604
frequency	lexicosyntactic	9.37E-09	-0.3725
info_window	semantic	1.27E-08	0.3420

Since the majority of the extracted acoustic features consist of MFCCs and measures related to aperiodicity of vocal fold vibration, the lack of significance of the acoustic features as discriminators between the two classes may be attributed to the fact that AD is not strongly associated with motor impairment of the articulators involved in speech production.

The second feature selection method is minimum-redundancy-maximum-relevance (mRMR), which minimizes the average mutual information between features and maximizes the mutual information between each feature and the class [18]; the features were ranked from most relevant to least. The results of this technique generally corroborate the selection made by the t-test, with no acoustic features among the top 10 selected. Here, mRMR selects a greater proportion of semantic features (e.g., mentions of the window and sink, and the number of occurrences of *curtain* and *stool*), placing more weight on the content of what the speaker is saying as a way of discriminating the two classes.

All of the features displayed in Table 2 have moderate statistically significant correlation with MMSE ($p < 0.001$). Since we are interested in the task of predicting clinical MMSE scores, the experiments described in Sec. 3 use correlation itself as a third feature selection method. The features are ranked by their correlation with MMSE, and the ones with the highest correlations are selected.

3. Experiments

3.1. Predicting MMSE score using LSAS features

To model the longitudinal progression of MMSE scores and LSAS features, we constructed a dynamic Bayes network (DBN) with continuous nodes, i.e., a Kalman filter with 2 variables, shown in Figure 1. Each time slice (Q_t, Y_t) represents one annual visit for a subject. Each conditioning node Q_t represents the underlying continuous MMSE score for that visit ($\mathbb{R}^{1\times1}$), while each node Y_t represents the vector of observed continuous LSAS features ($\mathbb{R}^{477\times1}$). A Kolmogorov-Smirnov test for normality was performed on the MMSE scores of all AD subjects, with the null hypothesis that they come from a normal distribution. The test did not reject this null hypothesis at the 5% confidence level, demonstrating that the data come from a

[2] Ongoing work distinguishes between AD and MCI.

Table 1: Summary of all extracted features (477 in total). The number of features in each type and subtype is shown in parentheses.

Type	Feature Subtype	Description and examples
Lexicosyntactic (182)	Production rule (121)	Number of times a production rule is used, divided by the total number of productions.
	Phrase type (9)	Phrase type proportion, rate and mean length.
	Syntactic complexity (4)	Depth of the syntactic parse tree.
	Subordination/coordination (3)	Proportion of subordinate and coordinate phrases to the total number of phrases, and ratio of subordinate to coordinate phrases.
	Word type (25)	Word type proportion; type-to-token ratio, Honoré's statistic.
	Word quality (10)	Imageability; age of acquisition (AoA); familiarity; transitivity.
	Length measures (5)	Average length of utterance, T-unit and clause, and total words per transcript.
	Perseveration (5)	Cosine distance between pairs of utterances within a transcript.
Acoustic (210)	MFCCs (170)	The first 42 MFCC parameters, along with their means, kurtosis and skewness, and the kurtosis and skewness of the mean of means.
	Pauses and fillers (8)	Total and mean duration of pauses; long and short pause counts; pause to word ratio; fillers (*um*, *uh*).
	Pitch and Formants (8)	Mean and variance of F0, F1, F2, F3.
	Aperiodicity (13)	Jitter, shimmer, recurrence rate, recurrence period density entropy, determinism, length of diagonal structures, laminarity.
	Other speech measures (11)	Total duration of speech, zero-crossing rate, autocorrelation, linear prediction coefficients, transitivity.
Sem. (85)	Mention of a concept (21)	Presence of mentions of indicator lemmas, related to key concepts in the Cookie Theft picture.
	Word frequency (64)	Number of times a given lemmatized word, relating to the Cookie Theft picture, was mentioned

normal distribution with M=18.52, SD=5.16. There are three conditional probability densities: the MMSE prior probability $P(Q_1)$, the MMSE transition probability $P(Q_t|Q_{t-1})$, and the LSAS feature observation probability $P(Y_t|Q_t)$.

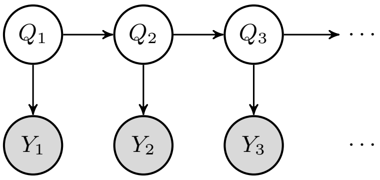


Figure 1: Temporal Bayes network (TBN) with continuous hidden (Q_t) and observed (Y_t) nodes. Hidden nodes represent MMSE score, and observed vectors represent LSAS features extracted from speech.

The feature set described in Sec. 2.2 is preprocessed to (i) remove features with zero variance across all samples, and (ii) normalize feature values to zero-mean and unit-variance, as is standard practice. Since the number of features (477) is large compared to the number of samples (393), the three feature selection methods described in Sec. 2.3 (i.e., a paired two-tailed t-test, mRMR, and correlation with MMSE score) are used to avoid overfitting, by varying the number of features selected by each method in order to determine the optimal feature set size.

The parameters of the three probability distributions in our model are trained using maximum likelihood estimation (MLE) since all training data are fully observed. During testing, the observed features for each test case are provided and junction tree inference on the trained model computes the marginal distribution of the now hidden (MMSE) nodes. Performance is measured as the mean absolute error (MAE) between actual and predicted MMSE scores. Since not all subjects have the same number of longitudinal samples, MAE is evaluated at the first and last hidden node, and averaged. Experiments are performed with leave-one-out cross-validation, where data from each subject, in turn, are used for testing and all other data for training, over all 255 subjects.

The results, with varying feature set sizes and feature selection methods, are shown in Table 3. The lowest MAE of 3.83 ($\sigma = 0.49$) is achieved when correlation is used to select the top 40 features. A two-factor repeated measures ANOVA performed on the mean MAE shows that both main effects are statistically significant, i.e., feature set size ($F_{7,24} = 8.67$, $p < 0.001$) and the feature selection method ($F_{2,24} = 4.07$, $p < 0.05$). The interaction effect is not significant ($F_{14,24} = 0.16$, ns), as expected given that the factors are independent.

To illustrate the longitudinal changes in cognitive and linguistic ability, Fig. 2 shows the pattern of decline of MMSE and the top 5 most correlated features for the subset of subjects with AD. This demonstrates the MMSE score declining non-monotonically over four annual visits (the maximum number of visits for AD subjects in DementiaBank), along with similar patterns across the indicated LSAS features.

3.2. Effect of longitudinal data on predicted MMSE score

To test the hypothesis that using longitudinal speech data aids in identifying underlying cognitive status (i.e., improving MMSE estimation), the Kalman filter experiment described in 3.1 is repeated for subsets of the dataset consisting of different amounts

Table 3: MAE in predicting MMSE scores using three feature selection methods and different feature set sizes. The lowest error for each feature selection method is highlighted in **bold**.

$N_{features}$	t-test	mRMR	ρ_{MMSE}
1	5.9788	5.3034	5.6396
5	5.6575	4.4440	5.0758
10	5.5148	4.3403	4.2098
20	5.2264	4.0426	4.1518
30	4.9066	4.1420	3.8628
40	**4.8073**	4.0648	**3.8333**
50	4.8520	**3.8551**	3.9180
all	7.3106	7.3106	7.3106

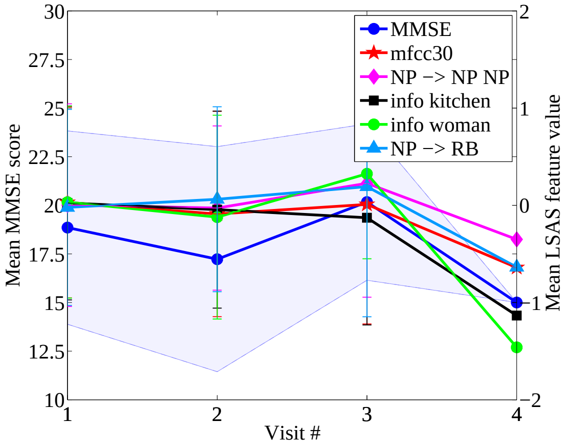


Figure 2: Pattern of decline of mean MMSE score and top 5 LSAS features most correlated with it, plotted versus annual visit number, for the subset of subjects with AD in DementiaBank. Standard deviation for MMSE is shown shaded in blue.

of longitudinal samples, T: (i) entire dataset (393 samples, 255 subjects, $1 \leq T \leq 5$), (ii) subset of subjects with 1 visit (154 samples, 154 subjects, $T = 1$), (iii) subset of subjects with at least two visits (239 samples, 101 subjects, $T \geq 2$), and (iv) subset of subjects with at least three visits (91 samples, 27 subjects, $T \geq 3$). The number of subjects with at least four visits is too low to conduct statistical experiments. The number of features used in the model is fixed to the optimal feature set size found in 3.1, and the feature selection method is varied (t-test, mRMR, correlation). Leave-one-out cross-validation is performed on each of the four datasets. The results are presented in Table 4. The lowest MAE for each feature selection method occurs on the dataset consisting of the highest number of longitudinal visits ($T \geq 3$). A two-factor repeated measures ANOVA performed on the mean MAE shows that the main effect of the data subset is statistically significant ($F_{3,9} = 5.43$, $p < 0.05$) while neither the second main effect ($F_{2,9} = 0.94$, ns) nor the interaction effect ($F_{6,9} = 0.54$, ns) is significant.

4. Discussion

Automatically extracted linguistic features can be used to effectively estimate underlying cognitive status, in terms of the most predominant clinical measure of dementia. The best result obtained with leave-one-out cross-validation on the entire dataset of 393 samples is an MAE of 3.83 ($\sigma = 0.49$), using correlation to select the top 40 features. This corresponds to a mean absolute relative error (MARE) of 21.0% (obtained as the absolute difference between predicted and actual MMSE score, divided by the actual MMSE score, and averaged over all runs). Molloy and Standish [19] reported that different rating styles among clinicians administering the MMSE and variance in test-retest scoring can lead to a within-subject interrater standard deviation of 3.9 to 4.8 and within-subject intrarater standard deviation of 4.8, with higher variation in low-scoring subgroups of subjects [1, 19]. The MAE obtained through statistical speech analysis in our present work is comparable to such variability. Further, the results obtained with the Kalman filter model significantly outperform an initial baseline multilinear regressor ran with leave-one-out cross-validation on the same dataset ($t = 2.31$, $p < 0.05$). This is being explored further.

Table 4: MAE in predicting MMSE score using three feature selection methods and different subsets of subjects with varied number of longitudinal datapoints. The lowest error for each feature selection method is highlighted in **bold**.

Dataset	t-test	mRMR	ρ_{MMSE}
all	4.807311	4.064823	3.8332502
1 visit	5.030811	4.978016	4.4916474
$\geq$ 2 visits	4.334934	3.534478	3.430414
$\geq$ 3 visits	**2.905163**	**3.063524**	**3.3577102**

The fact that correlation outperforms the other two feature selection methods is expected, as it computes the relationship between the features and the MMSE score directly whereas the others use the presumed diagnosis to dichotomize the data into classes. The majority of features selected on each iteration of cross-validation are typically lexicosyntactic and semantic, with acoustic features typically not being among the most relevant. While this may suggest that anatomical irregularities in speech production are less meaningful, we note that the lexicosyntactic features depend, to a large extent, on the free expression of language through speech. Specifically, the working memory impairment associated with AD affects preferred syntactic constructions in speech, leading to shorter utterances, fewer complex noun and verb phrases, a higher number of pronouns, and lexical impoverishment indicated by Honoré's statistic.

We also show that focussing on subsets of subjects with a higher number of longitudinal samples improves the accuracy of inference in the Kalman filter model, lowering MAE to 2.91 ($\sigma = 0.31$) or equivalently lowering MARE to 12.5%, using a t-test for selecting the top 40 features. Since DementiaBank contains a variable number of samples for each subject, the number of subjects and the proportion of subjects with AD in each subgroup explored in Sec. 3.2 is not balanced. We therefore suggest that future data collection of pathological speech should involve more longitudinal samples across participants.

While MMSE is one of the most widely used clinical tests for cognitive ability, it is somewhat coarse, lacking sensitivity to subtle changes in cognition in the early stages of dementia, as well as having a high false-negative rate in addition to inter-annotator disagreement and test-retest variability [20, 1, 19]. While automated prediction of the MMSE score may aid the screening process for AD by reducing the cost and time involved, and improving reliability, future work will explore more precise measures of cognitive decline. The Montreal Cognitive Assesent (MoCA) and the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) [21] are screening tests which have been shown to have higher sensitivity than

MMSE to subtle changes in cognitive decline in populations with MCI and mild dementia [22]; future studies are needed to assess the validity of automatic scoring of such tests as a more fine-grained measure of the progression of cognitive decline.

5. Acknowledgements

This work is funded by an NSERC Discovery grant (RGPIN 435874) and by a Young Investigator award by the Alzheimer Society of Canada.

6. References

[1] D. W. Molloy, M. B. E. Alemayehu, and R. Roberts, "Reliability of a Standardized Mini-Mental State Examination compared with the traditional Mini-Mental State Examination," *The American Journal of Psychiatry*, vol. 148, no. 1, pp. 102–105, 1991.

[2] C. Ballard, S. Gauthier, A. Corbett, C. Brayne, D. Aarsland, and E. Jones, "Alzheimer's disease," *The Lancet*, vol. 377, no. 9770, pp. 1019–1031, 2011.

[3] R. M. Li, A. C. Iadarola, and C. C. Maisano. (2007) Why population aging matters: A global perspective. [Online]. Available: http://www.nia.nih.gov/research/publication/why-population-aging-matters-global-perspective

[4] R. Sperling, P. Aisen, L. Beckett, D. Bennett, S. Craft, A. Fagan, T. Iwatsubo, C. R. J. Jr., J. Kaye, T. Montine, D. Park, E. Reiman, C. C. Rowe, E. Siemers, Y. Stern, K. Yaffe, M. C. Carrillo, B. Thies, M. Morrison-Bogorad, M. V. Wagster, and C. H. Phelps, "Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging—Alzheimer's Association workgroups on diagnostic guidelines," *Alzheimer's and Dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[5] M. Folstein, S. E. Folstein, and P. R. McHugh, "Mini-Mental State: a practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[6] S. Kemper, M. Thomas, and J. Marquis, "Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content," *Psychology and Aging*, vol. 16, no. 4, pp. 600–614, 2001.

[7] S. Orimaye, J. S.-M. Wang, and K. J. Golden, "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proc. of the ACL 2014 Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Baltimore, USA, Jun. 2014, pp. 78–87.

[8] H. Goodglass and E. Kaplan, *The Assessment of Aphasia and Related Disorders*. Philadelphia, PA: Lea and Febiger, 1983.

[9] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proc. of the ACL 2014 Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Baltimore, USA, 2014, pp. 27–37.

[10] X. Le, I. Lancashire, G. Hirst, and R. Jokel, "Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists," *Literary and Linguistic Computing*, vol. 26, no. 4, pp. 435–461, 2011.

[11] M. Sundermann, "Longitudinal effects of Alzheimer's disease," Master's thesis, Department of Speech and Hearing Science, Ohio State University, 2012.

[12] H. Bird, M. A. L. Ralph, K. Patterson, and J. R. Hodges, "The rise and fall of frequency and imageability: Noun and verb production in semantic dementia," *Brain and Language*, vol. 73, pp. 17–49, 2000.

[13] K. Gilhooly and R. Logie, "Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words," *Behavior Research Methods*, vol. 12, pp. 395–428, 1980.

[14] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proc. of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430.

[15] J. O. de Lira, K. Z. Ortiz, A. C. Campanha, P. H. F. Bertolucci, and T. S. C. Minett, "Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease," *International Psychogeriatrics*, vol. 23, no. 3, pp. 404–412, 2011.

[16] H. Stadthagen-Gonzalez and C. J. Davis, "The Bristol norms for age of acquisition, imageability, and familiarity," *Behavior Research Methods*, vol. 38, no. 4, pp. 598–605, 2006.

[17] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans Biomed Eng.*, vol. 59, no. 5, pp. 1264–1271, 2011.

[18] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 27, no. 8, pp. 1226–1238, 2005.

[19] D. W. Molloy and T. I. M. Standish, "A guide to the Standardized Mini-Mental State Examination," *International Psychogeriatrics*, vol. 9, no. 1, pp. 87–94, 1997.

[20] A. Nelson, B. Fogel, and D. Faust, "Bedside cognitive screening instruments — a critical assessment," *The Journal of Nervous and Mental Disease*, vol. 174, no. 2, pp. 73–83, 1986.

[21] C. Randolph, *Repeatable Battery for the Assessment of Neuropsychological Status Update*. San Antonio, TX: The Psychological Corporation, 2012.

[22] C. Zadikoff, S. Fox, D. Tang-Wai, T. Thomsen, R. de Bie, P. Wadia, J. Miyasaki, S. Duff-Canning, A. Lang, and C. Marras, "A comparison of the Mini Mental State Exam to the Montreal Cognitive Assessment in identifying cognitive deficits in Parkinson's disease," *Movement Disorders*, vol. 23, no. 2, pp. 297–299, 2008.

[23] D. Bone, T. Chaspari, K. Audkhasi, J. Gibson, A. Tsiartas, M. V. Segbroeck, M. Li, S. Lee, and S. Narayanan, "Classifying language-related developmental disorders from speech cues: the promise and the potential confounds," in *Proc. of the 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 182–186.

[24] L. Meteyard and K. Patterson, “The relation between content and structure in language production: an analysis of speech errors in semantic dementia,” *Brain and Language*, vol. 110, no. 3, pp. 121–134, 2009.

[25] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, “Spoken language derived measures for detecting mild cognitive impairment,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, 2011.

[26] D. G. Silva, L. C. Oliveira, and M. Andrea, “Jitter estimation algorithms for detection of pathological voices,” *EURASIP Journal on Advances in Speech Processing*, vol. 2009, pp. 1–9, 2009.

From European Portuguese to Portuguese Sign Language

Inês Almeida[1], *Luísa Coheur*[1], *Sara Candeias*[2]

[1]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa
[2]Microsoft Language Development Center, Lisbon, Portugal

name.surname@L2F.inesc-id.pt, t-sacand@microsoft.com

Abstract

Several efforts have been done towards the development of platforms that allow the translation of specific sign languages to the correspondent spoken language (and vice-versa). In this (demo) paper, we describe a freely available system that translates, in real time, European Portuguese (text) into Portuguese Sign Language (LGP), by using an avatar. We focus in how some linguistic elements are dealt in LGP, and in the Natural Language Processing (NLP) step implemented in our system. The system's interface will be used to demonstrate it. Although only a small set of linguistic phenomena are implemented, it can be seen how the system copes with it.

Index Terms: sign languages, translation, natural language processing, portuguese, avatar

1. Introduction

Sign languages are now considered full natural human languages. Contrary to auditory-vocal languages, sign languages are visual-gestural languages that merge manual communication and body language [1]. The meaning is expressed with a combination of different hand shapes, orientation and movement of the hands (manual features). Non-manual features, such as body movements (upper torso) and facial expressions are also used, as well as fingerspelling – the process of spelling out words by using signs that correspond to the letters of the word in the local vocal language.

Sign Languages have their own vocabulary and grammatical rules, which do not match the correspondent spoken language as the writing system does [1]. For instance, American Sign Language and British Sign Language are different and not mutually understandable, although learnt by people living in English speaking countries. Thus, it is very difficult to take advantage of existing resources when moving to a new sign language.

For some languages, several studies emerged in the last years, with their focus ranging from linguistic and humanistic to automatic translation. However, only very recently, LGP (officially recognised in 1997[1]) has been a target of these studies. Currently, there are several dictionaries [2, 3] both in image and video format, but only one grammar [4] in a very incomplete state. There is no official number for deaf persons in Portugal, but the 2011 census [5] mentions 27,659 deaf persons, making, however, no distinction in the level of deafness, and on the respective level of Portuguese and LGP literacy. Aiming to contribute to this community, we developed a system, which, given as input a sentence in (European) Portuguese, performs the correspondent signs in LGP, by using an avatar. At the basis of this system, there is a flexible architecture that takes advantage of NLP tools, as these can give an important contribution to the translation process. For instance, if a proper noun is identified, if no sign is associated with it, fingerspelling is the solution. Moreover, as we will see, in some cases, a word can be signed by signs associated with its root and suffixes. Thus, a stemmer or a Part-of-Speech (POS)-tagger can play a fundamental role in these situations. A detailed description of the system can be found in [6] and [7]. The system can be downloaded from `http://web.ist.utl.pt/~ist163556/pt2lgp`.

This paper is organised as follows: in Section 2 we present related work, in Section 3 we describe some basic linguistic phenomena in LGP, in Section 4 we describe our prototype, and, in Section 5, we explain what can be tested in our demo. Finally, in Section 6 we conclude and point to future work.

2. Related Work

Many efforts were done towards the development of translators from different sign languages to their spoken counterparts and vice-versa. A number of projects in the area are focusing in the entire system pipeline (from spoken to sign languages and vice-versa), as the work of [8], for Portuguese, and [9] for Mandarin; others only target part of it (for instance [10], which deals with Italian Sign Language). Current trends in Automatic Machine Translation cannot be followed as there are no parallel corpora (except in some specific contexts) to train the translators. Thus, most of the existing systems are based on handcrafted glosses, relating signs with words, which is also our approach.

Recently LGP has been the focus of several computational studies. The work described in [11] focus on avatars, and on how to produce avatars signs, based on human signs; the work in [12] targets the teaching of LGP; the Virtual Sign Translator [8] contributes with a translator between European Portuguese and LGP, and it was also applied to be used in a game that teaches LGP [13]. However, to the best of our knowledge, none of these works explored how current NLP tasks can be applied to help the translation process.

3. Linguistic Concepts

In this section we make a brief overview of some linguistic phenomena in LGP. At the basis of our study are the static images of hand configurations presented in an LGP dictionary [2], LGP videos from different sources, such as the Spread the Sign initiative[2], and, the (only) LGP grammar [4], from 1994. LGP interpreters were also consulted, as we could not found information regarding some linguistic phenomena in the previous mentioned materials.

[1] `http://www.fpasurdos.pt/legislacao/decretos-e-leis/`

[2] `http://www.spreadthesign.com`

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 140–143,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

3.1. Nouns

Concepts in LGP usually do not have an associated gender, and, thus, do not need inflection. For animated beings, and when relevant, gender can be specified with a prefix, expressing ‘male’ or ‘female’ (as for ‘coelha’ (‘female rabbit’), which becomes ‘female’ + ‘rabbit’). In case of omission, the male gender is assumed. However, there are classes of nouns that are female by default as is the case of ‘enfermeira’ (‘nurse’), and need to be obligatorily prefixed with ‘male’. Another (more common) exception is to have two separate words to denote the male and female case, as in ‘leão’ (‘lion’) and ‘leoa’ (‘lioness’).

Considering plural cases for LGP, this can be done in several different ways. The first is *incorporation*, allowing to explicitly specify a small quantity after the noun. Examples are ‘pessoas+4’ (‘persons+4’), or ‘pessoas+muitas’ (‘persons+many’). The second is *repetition*, meaning to perform a sign multiple times as seen for ‘árvores’ (trees). The last is *reduplication*, meaning to make the sign with both hands as in ‘pessoas’ (persons). However, there are many non identified processes for LGP and the cases of the usage of each type of plural are not clear.

With regard to proper nouns, fingerspelling is often used. If the person does not have a known gestural name, fingerspelling the letters of her/his name is the solution.

3.2. Adjectives

The sign for the adjective follows the sign for the noun. Figure 1 illustrates the signs for ‘coelho pequeno’ (‘little rabbit’).

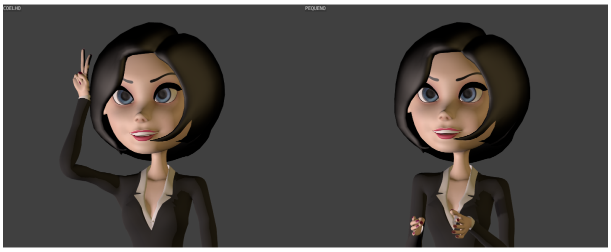


Figure 1: Signs for ‘coelho pequeno’

Notice that, if the signs for ‘coelho’ and ‘pequeno’ are available (and although this cannot be seen as a rule) words as ‘coelhinho’ (also ‘little rabbit’), can also be translated, as long as we are able to properly identify suffixes.

3.3. Numbers

Numbers can be used as a quantitative qualifier, isolated number (cardinal), ordinal number, and composed number (*e.g.* 147). Signs associated with each number also vary their forms if we are expressing a quantity, a repetition or a duration, and if we are using them as an adjective or complement to a noun or verb. Reducing the test case to ordinal numbers, the main difficulty is to express numbers in the order of the tens and up. For instance, ‘147’ is signed as ‘1’, followed by ‘4’ and ‘7’ with a slight offset in space as the number grows. Numbers from ‘11’ to ‘19’ can be signed with a blinking movement of the units’ number. Some numbers, in addition to this system, have a totally different sign (as *e.g.* ‘11’, which has its own sign).

3.4. Verbs

When the use of the verb is plain, with no past or future participles, the infinitive form is used in LGP. Most verbs are inflected according to the associated subjects and are affected by the action, the time and the way the action is realised. For instance, for the regular use of the verb ‘to eat’, the hand goes twice to the mouth, closing from a relaxed form, with palm up. However, this verb in LGP is highly contextualised with what is being eaten. Thus, the verb should be signed recurring to different hand configurations and expressiveness, describing *how* the thing is being eaten (not all the deaf associations agree on this).

The Portuguese grammar [4] refers a temporal line in the gesturing space with which verbs should concord with in past, present and future tenses. The verb inflection is made along this imaginary line with eye, eyebrow and upper body movement. A common practice is to add a time adverb to the sentence, such as *passado* ‘past’, *futuro* ‘future’ or *amanhã* ‘tomorrow’. The adverbial expression is also performed along the timeline with a possible emphasis on the distance in time. For example, the word *agora* ‘now’ is always signed in front of the signer and close to the torso, but it can be signed more and more close to express immediateness or the reverse to express laxness. This is a feature often found in other sign languages.

In what concerns verb agreement, to the best of our knowledge, there is no gender or number agreement in LGP. This information must be express by direct referencing to the subject, for example, by mentioning a personal pronoun before the verb. For instance, in the sentence *eu pergunto-te* ‘I ask you’, the verb is directed from ’I’ to ’you, while in the sentence *tu perguntas-me* ‘You ask me’, the verb changes directionality. Additionally, the pronoun ’you’ is signed in the direction of the second person’s face in the case of the verb ‘ask’, but in the direction of the chest with the verb ‘to give’.

Modality is realised throughout the imaginary temporal line, indicating duration and repetition through movement. An example is the verb *andar* ‘to walk’, which is signed with different movement modulation for *andar* ‘walk’, *ir andando* ‘walking’, *andar apressadamente* ‘walk hurriedly’, *andar pesadamente* ‘stumping’ and so on.

3.5. Syntax

Syntax in sign languages is made by spatial agreement of signs. To the best of our knowledge, there are no studies at the sentence level for the LGP, but studies for American Sign Language (ASL) [14], indicate the existence of several complex phenomena, such as *loci* and *surrogates* for the agreement of verbs with virtual entities. However, it is known that in a syntactic point of view, LGP is Object–Subject–Verb (OSV), while spoken Portuguese is predominantly Subject–Verb–Object (SVO).

4. The prototype

The Natural Language ToolKit (NLTK)[3] was used in all the NLP tasks. Blender[4] was our choice regarding the 3D package for animation. Both are widely used, community driven, free and open source. Moreover, NLTK offers taggers, parsers, and other tools in several languages, including Portuguese. In the following we describe each one of the main tasks of our system. Figure 2 presents the general architecture of our prototype.

[3] http://www.nltk.org

[4] http://www.blender.org

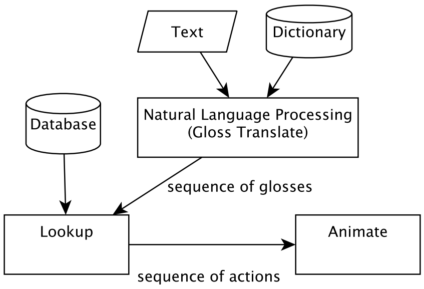


Figure 2: Proposed architecture

4.1. Natural Language Processing

Several well established tasks in the NLP field where integrated in our system, namely:

- **Error correcting and normalization**: A step that enforces lowercase and the use of latin characters. Common spelling mistakes can be corrected in this step.
- **Tokenizer**: The input string is split into sentences and then into words. The tokenizer, provided by NLTK, uses Portuguese language training. Example of a tokenized input: ['o', 'joão', 'come', 'a', 'sopa'].
- **Stemmer**: As a form of morphologic parsing, we apply a stemmer that identifies suffixes and prefixes to use as an adjective or classifier to the gloss. This allows, for example, 'coelhinha' ('little female rabbit'), to be understood, by its suffixes ('inho' +'a') , to be a small ('inho) and a female ('a') derivation of the root 'coelh(o)'.
- **POS-Tagger**: We make use of NLTK's n-gram taggers, starting with a bigram tagger, with a backoff technique for an unigram tagger and the default classification of 'noun' (the most common class for Portuguese). We used the treebank 'floresta sintá(c)tica' corpus [15] for training the taggers. Using the same example, the result would be: [('o', 'art'), ('joão', 'prop'), ('come', 'v-fin'), ('a', 'prp'), ('sopa', 'n')].
- **Named Entity Extraction**: We apply Named Entity Recognition (NER) for identifying names of persons, by matching against a list of common Portuguese names.
- **Lexical Transfer**: The expanded and annotated list of words are converted to their corresponding glosses using a dictionary lookup. This results in items such as ['GLOSS', ['SOPA']] and ['FINGERSPELL', ['J', 'O', 'A', 'O']].
- **Structure Transfer**: The prototype supports reordering of adjectives and quantities to the end of the affecting noun, for example the input *dois coelhos* ('two rabbits') would result in [['GLOSS', ['COELHO']], ['NUMERAL', ['2']] ('coelho + 2'). The prototype also supports basic reordering of sequences of 'noun - verb - noun', in an attempt to convert the SVO ordering used in Portuguese to the more common structure of OSV used in LGP.

4.2. Lookup

The animation lookup, given a gloss, is done via a JSON file mimicking a database. The database consists of a set of *glosses* and a set of *actions*. The action ids are mapped to blender actions, that are in turn referenced by the glosses. One gloss may link to more than one action, that are assumed to be played sequentially.

4.3. Animation

We implemented base hand configurations. These differ from sign language to sign language. For LGP there are 57 base configurations, composed of 26 hand configurations for letters, 10 for numbers, 13 for named configurations and 8 extra ones matching greek letters (examples in Figure 3).

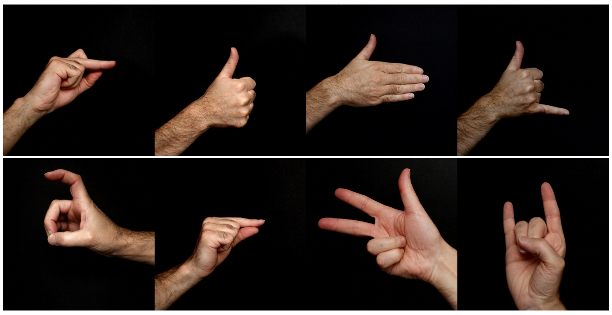


Figure 3: Some hand configurations in LGP.

Doing a set of base hand configurations to start, proved to be a good choice as it allowed to test the hand rig and basic methodology. All the 57 basic hand configuration were manually posed and keyed from references gathered from [2, 4, 3], and also from the Spread the Sign project videos[5]. The ten (0 to 9) implemented hand configurations are shown in Figure 4.

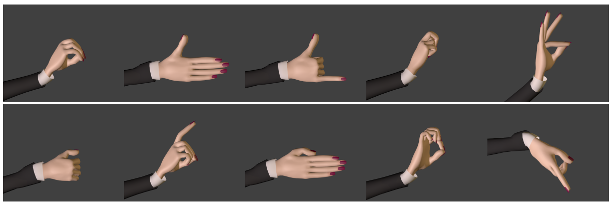


Figure 4: Hand configurations for numbers (0-9)

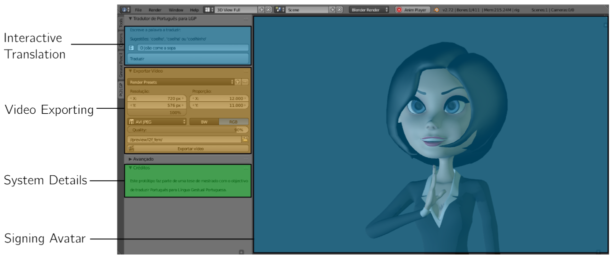
The animation is synthesised by directly accessing and modifying the action and f-curve data. We always start and end a sentence with the rest pose. For concatenating the actions, we *blend* from one to the other in a given amount of frames by using Blender's Non-Linear Action tools that allow action layering. Channels that are not used in the next gesture, are blended with the rest pose instead. We adjust the number of frames for blending according to the hints received. For fingerspelling mode, we expand the duration of the hand configuration (which is originally just one frame). Further details about this process can be found in [6] and [7].

5. The demo

Users can interact with our system via an interface, which consists of an input text box, a button to translate, and a 3D view with the signing avatar. The 3D view can be rotated and zoomed, allowing to see the avatar from different perspectives.

The breakdown down in Figure 5 shows the interface.

[5] http://www.spreadthesign.com



v2.72 | Bones:1/411 | Mem:215.24M | rig Scenes:1 | Cameras:0/0

▼ Tradutor de Português para LGP

Escreve a palavra a traduzir:

Sugestões: 'coelho', 'coelha' ou 'coelhinho'

O João come a sopa

▼ Exportar Vídeo

Resolução: X: 720 px Y: 576 px 100%

Proporção: X: 12.000 Y: 11.000

AVI JPEG

Quality: 90%

//preview/l2f_fem/

▶ Avançado

▼ Créditos

Este protótipo faz parte de uma tese de mestrado com o objectivo de traduzir Português para Língua Gestual Portuguesa.

Figure 5: User Interface for the prototype

Additionally, we provide an interface for exporting video of the signing that supports choosing the resolution, aspect ratio and file format. This panel is indicated in the image in orange. Displayed in green, is a panel indicating the authors and describing the project. All panels but the one used for the main interaction start folded. It should be clear that it is still possible to use extra functionalities of Blender, thus making advanced usage of the system.

In what concerns current possibilities of the system, common spelling mistakes in the words used for the test cases can be corrected. Moreover, several words deriving from the stem ‘coelho’ were implemented, such as ‘coelha’ (female rabbit) and ‘coelhinho’ (little rabbit). Besides isolated words, some full sentences, such as ‘O João come a sopa’, can be tested. The verb sign had to be extended, as for eating soup, it is done as if handling a spoon (for instance, for eating apples, the verb is signed as if holding the fruit).

To conclude, we should say that two deaf associations were reached for a preliminary evaluation. Feedback on clarity and readability was very positive.

6. Conclusions and Future Work

In this paper we described the system we would like to demonstrate, focusing on its NLP component. It was designed to be free and open-source. All the basic hand signs for LGP were implemented, as well as the whole basic infrastructure (already accommodating different language phenomena).

This work led to a collaborative project between academia and industry that aims at creating a prototype that translates European Portuguese (text and speech) into LGP, in real time. As future work, besides moving to the translation between LGP and European Portuguese, we will extend the database and the dictionaries. Also, we will work in an interface that will allows us to easily add data to the system. Current NLP tasks and techniques will also be further explored. A more formal evaluation also needs to be done.

7. Acknowledgements

We would like to thank to Associação Portuguesa de Surdos and Associação Cultural de Surdos da Amadora for all their help. However, the responsibility for any imprecision lies with the authors alone. This work was partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013. Microsoft Language Development Center is carrying this work out in the scope of a Marie Curie Action IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP).

8. References

[1] H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” in *Visual Analysis of Humans: Looking at People*, T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, Eds. Springer, Oct. 2011, ch. 27, pp. 539 – 562.

[2] A. B. Baltazar, *Dicionário de Língua Gestual Portuguesa*. Porto Editora, 2010.

[3] A. Ferreira, *Gestuário: língua gestual portuguesa*. SNR, 1997.

[4] M. Amaral, A. Coutinho, and M. Martins, *Para uma gramática da Língua Gestual Portuguesa*, ser. Colecção universitária. Caminho, 1994. [Online]. Available: http://books.google.pt/books?id=yZ2PQAAACAAJ

[5] Instituto Nacional de Estatística (INE), “Census 2011, xv recenceamento geral da população, v recenceamento geral da habitação, resultados definitivos – portugal,” INE, Tech. Rep., 2012.

[6] I. R. Almeida, “Exploring challenges in avatar-based translation from european portuguese to portuguese sign language,” Master’s thesis, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, 2104.

[7] I. Almeida, L. Coheur, and S. Candeias, “Coupling natural language processing and animation synthesis in portuguese sign language translation,” in *Vision and Language 2015 (VL15), EMNLP 2015 workshop (accepted for publication)*, Lisbon, Portugal, 2015.

[8] P. Escudeiro, N. Escudeiro, R. Reis, M. Barbosa, J. Bidarra, A. B. Baltazar, and B. Gouveia, “Virtual sign translator,” in *International Conference on Computer, Networks and Communication Engineering (ICCNCE)*, A. Press, Ed., Chine, 2013.

[9] X. Chai, G. Li, X. Chen, M. Zhou, G. Wu, and H. Li, “Visualcomm: A tool to support communication between deaf and hearing persons with the kinect,” in *ASSETS 13: Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, NY, USA: ACM, 2013.

[10] D. Barberis, N. Garazzino, P. Prinetto, and G. Tiotto, “Improving accessibility for deaf people: An editor for computer assisted translation through virtual avatars,” in *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS ’11. New York, NY, USA: ACM, 2011, pp. 253–254.

[11] J. Bento, “Avatares em língua gestual portuguesa,” Master’s thesis, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal, 2103.

[12] J. Gameiro, T. Cardoso, and Y. Rybarczyk, “Kinect-sign, teaching sign language to listeners through a game,” *Procedia Technology*, vol. 17, no. 0, pp. 384 – 391, 2014.

[13] P. Escudeiro, N. Escudeiro, R. Reis, M. Barbosa, J. Bidarra, A. B. Baltasar, P. Rodrigues, J. Lopes, and M. Norberto, “Virtual sign game learning sign language,” in *Computers and Technology in Modern Education*, ser. Proceedings of the 5th International Conference on Education and Educational technologies, Malaysia, 2014.

[14] S. K. Liddell, *Grammar, gesture, and meaning in American Sign Language*. Cambridge: Cambridge University Press, 2003.

[15] S. Afonso, E. Bick, R. Haber, and D. Santos, “Floresta Sintáctica: A treebank for Portuguese.” *LREC*, pp. 1698–1703, 2002. [Online]. Available: http://beta.visl.sdu.dk/pdf/AfonsoetalLREC2002.ps.pdf