Analysis of phonetic transcriptions for Danish automatic speech recognition

Andreas Søeborg Kirkedal

Department of International Business Communication, CBS
Dalgas Have 15, DK-2000 Frederiksberg
Denmark

`ask.ibc@cbs.dk`

Abstract

Automatic speech recognition (ASR) relies on three resources: audio, orthographic transcriptions and a pronunciation dictionary. The dictionary or lexicon maps orthographic words to sequences of phones or phonemes that represent the pronunciation of the corresponding word. The quality of a speech recognition system depends heavily on the dictionary and the transcriptions therein. This paper presents an analysis of phonetic/phonemic features that are salient for current Danish ASR systems. This preliminary study consists of a series of experiments using an ASR system trained on the DK-PAROLE corpus. The analysis indicates that transcribing e.g. stress or vowel duration has a negative impact on performance. The best performance is obtained with coarse phonetic annotation and improves performance 1% word error rate and 3.8% sentence error rate.

Keywords: Automatic speech recognition, phonetics, phonology, speech, phonetic transcription.

1 Introduction

Automatic speech recognition systems are seeing wider commercial use now more than ever before. No longer are ASR systems restricted to rerouting scenarios in call centres with a small and domain-specific vocabulary. The largest commercial experiment in Europe to date has taken place in the Danish public sector in Odense municipality and entailed more than 500 case workers dictating reports rather than typing them.

To be a practical alternative to manual transcriptions or typing in general, the recognition rate must be high. Otherwise, the potential gain will be spent correcting the recognised output. Many approaches to optimisation of ASR have been investigated including multiple pronunciation variants, direct modelling of acoustic features and domain-specific language modelling. One thing that seems to be missing is an investigation of what features of pronunciation are salient for the ASR system. This is likely a result of the complexity of ASR systems, meaning that ASR research is usually conducted by computer scientists or engineers, who can better understand the intricacies of acoustic models, language models, acoustic feature extraction etc.

1.1 Related work

From a phonetic point of view, research into the importance of the transcription chosen for the pronunciation dictionary and the phonetic features in the alphabet has not been carried out extensively. The phonetic transcriptions available in corpora are usually not created for computational modelling. In many cases, the description *itself* is the goal rather than what the description can be used *for*. Therefore, the transcription contains symbols for phonetics and prosody.

One related study investigates the effect of rate of speech (ROS) on Danish ASR (Brøndsted and Madsen, 1997). The article shows that low and high ROS have a negative impact on recognition accuracy and that e.g. long and short vowels should be modelled separately because long vowels are more sensitive to ROS than short vowels are.

Modelling long and short vowels separately supports the observation that all vowels in Danish have a short and a long version and this a distinctive lexical feature (Grønnum, 2005).

(Ljolje, 1994) introduced the use of Gaussian Mixture Models (GMMs) and probabilistic duration modelling to distinguish between long and short vowels. Using duration modelling and GMMs the word error rate (WER) on a 25000 word task was reduced by 10% and GMMs have been part of standard ASR systems ever since.

Schwa assimilation in Danish has been investigated in (Schachtenhaufen, 2010, In Danish). From a linguistic point of view the article describes a set of rules for schwa assimilation, schwa elision and distinct schwa pronunciation based in part on sonority principles and on syllable and word boundaries. This is further condensed into a single rule for the manifestation of schwa assimilation. The linguistic investigation is conducted on the DanPASS corpus and is mainly a qualitative study.

2 Speech recognition overview

An ASR system consists of an acoustic model, a language model and a pronunciation dictionary. The acoustic model is a phone classifier. The input for an acoustic model is a vector of acoustic parameters such as Mel Feature Cepstral Coefficients (MFCC). The feature vector is extracted from 10-20 ms windows of the speech input. Each of the MFCC vectors are then classified as

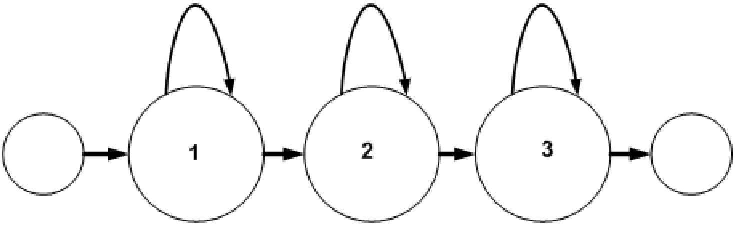

Figure 1: A triphone model.

a set of phones with an associated probability. Usually, there is a n-to-1 relationship between MFCC vectors and phonetic symbols, because the phone duration will be longer than the window a vector is derived from.

To take coarticulation effects into account and make the phone classification more context-dependent, each phone is subdivided into three subphones: *start*, *middle* and *end* as shown in Figure 1. The *start* state is the transition into the current phone and is dependent on the previous phone, the *middle* state is the current phone, which is most context-independent and the *end* state is the transition out of this phone, which is sensitive to the next phone and affects the *start* of the next phone. These groups of subphones are called *triphones*. An n-to-1 relationship holds between MFCC vectors and subphones as well and is handled in a triphone model by allowing transitions from a state back to the same state.

Sequences of triphones are compared to the pronunciation dictionary and a set of possible orthographic sentences or phrases are created from the triphone sequences. They are then weighted by the language model and ranked according to the acoustic model score and the language model score. The highest scoring sentence is found using e.g. the Viterbi algorithm and output by the decoder as the result. A toy example of the decoding process can be seen in Figure 2.

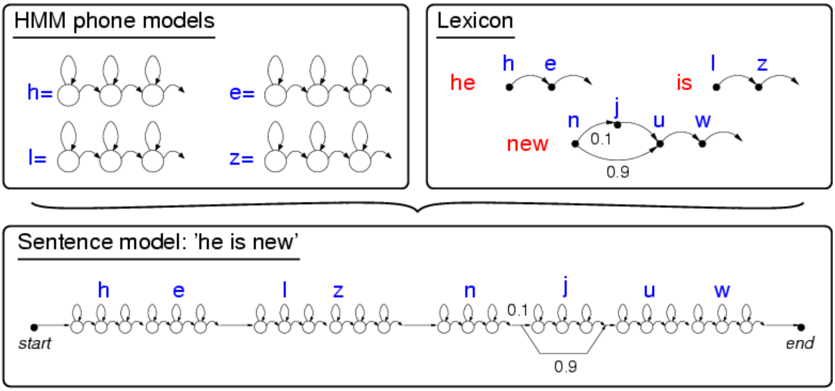

Figure 2: A toy decoding example. Notice that in this example, a pronunciation variant of *new* is allowed.

2.1 Training

ASR is data-driven and needs large amounts of data to train acoustic and language models. The type of data is important for the performance of the ASR system and features such as voice, domain, noise, volume and variation in the training data has an impact on the performance of an ASR system. Recordings must be transcribed and a phonetic dictionary with a transcription for each word must be created or acquired. The audio files and the transcriptions are fundamental to train an ASR system. The audio files must be recorded in such a way that they are suitable for training an ASR system. The audio files should preferably be in uncompressed format such as MSWAV, a sampling rate of 16kHz, 16 bit and use only a single channel, i.e. mono, not stereo. To train a recogniser for telephone speech, the sampling rate must be 8kHz. If you have audio files, which have a higher sampling rate and/or is in stereo, it is possible to downsample and convert into mono, but it is not possible to upsample your data.

3 Phonetic transcription

Several alphabets are used to create phonetic or phonemic transcriptions. They have been used by researchers for decades and can be seen in almost any dictionary entry. Depending on the researcher or the purpose of a transcription, the full expressive power of an alphabet can be used or a subset thereof. Common for all the alphabets mentioned below are that they can also express suprasegmentals such as stress or Danish *stød*.

3.1 International Phonetic Alphabet

Most known is IPA (International Phonetic Association, 1999), which is used extensively around the world by researchers in and teachers of phonetics and phonology. The IPA alphabet contains symbols for the pronunciation of vowels and consonants as well as symbols or diacritics for different positions of speech organs, duration and prosodic features such as stress and tones.

3.2 Speech Assesment Methods Phonetic Alphabet

Speech Assesment Methods Phonetic Alphabet (Wells et al., 1997) (SAMPA) is a machine-readable language specific phonetic alfabet. It is developed as a collaborative effort between native speakers and phoneticians of all the languages it can be applied to. An extension - X-SAMPA - is a recoding of IPA symbols in ASCII symbols to make the symbols machine-readable.

4 Experiment

The goal is to study the impact of transcription granularity and different phone sets on word error rate (WER). To be able to conduct these experiments, a corpus or corpora suitable for training an ASR system that contains high quality transcription is needed. In addition, a strategy for converting phonetic/phonemic transcriptions to a machine-readable format and for filtering the transcriptions must be chosen and finally, a toolkit for training the ASR systems must be chosen.

4.1 Corpora

There are several Danish spoken language corpora. Some suitable corpora are listed here:

- LANCHART
- DanPASS

- DK-Parole

LANCHART (Gregersen, 2007) contains more than 1000 hours of spontaneous speech. The corpus is automatically transcribed, which means the transcription is generated from orthographic transcription, and all annotations are encoded in PRAAT textgrids (Boersma, 2002). The corpus is designed to monitor language change over time and is representative with respect to regional dialects, age, gender, educational background etc. However, due to the highly overlapping nature of the interviews, the recordings are not suitable for training ASR systems. It is likely that a subcorpus would be suitable, but this kind of filtering is beyond the scope of the project due to the size of the original corpus.

DanPASS (Grønnum, 2006) is a highly accurate, manually annotated speech corpus with time-coded phonetic and phonemic transcriptions, POS, lemmata etc. It consists of two subcorpora containing monologues and dialogues. The dialogues corpus contains recordings of experiments with two participants. Each participant has a map and one participant must guide the other to follow a route on the map. However, the two maps are not identical. When the participants must explain to each other the differences in the maps, semi-spontaneous speech occurs during the unscripted negotiation of the different maps. The monologues corpus contains recordings where one participant is asked to describe a network of coloured geometric networks. Even though the corpus contains dialogues, there are no overlapping speech. Speakers of different age, gender and different dialects are represented. The annotation in the corpus is the most fine-grained in all the corpora using most of the expressive power of IPA. DanPASS also uses textgrids to encode annotations and transcriptions.

DK-Parole (Henrichsen, 2007) contain recordings of read-aloud speech of one speaker. The annotation is not as fine-grained as DanPASS, but uses the same alphabet as LANCHART, is manually annotated and easier to use for training data. In addition, the SAMPA alphabet is created to be machine-readable and therefore does not require any conversion of the transcriptions to be usable for training an ASR system.

4.2 Phonetic character conversion

Character encoding conversion is possible with existing methods. In the case of IPA symbols, the character encoding must be changed to ASCII from Unicode and the best resource is conversion with X-SAMPA[1]. In addition, ASR toolkits have reserved symbols such as underscore ("_") because they are used for e.g. segmentation during computation that also must be filtered in a way that does not introduce unintended ambiguity. The symbols will be usable for the machine learning algorithms, but may become unintelligible to linguists, which constitutes a problem for debugging.

4.3 Transcription filtering

The phonetic and/or phonemic transcriptions in all the corpora from Section 4.1 contain symbols and diacritics for segments and suprasegments. The classical phonological definition of segments is the smallest discrete unit of speech that you can find (Grønnum, 2005). All the phones in Figure 2 are segments. Suprasegments are phenomena in speech that cover a larger time domain than segments. In general, suprasegmental features such as tone, stress, stød and duration are related to prosody, which is related to syllables, words and phrases.

[1] See `http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm`.

Suprasegmental features are annotated with diacritics[2].

In current ASR systems, there are no concepts of syllables, stress groups or prosody in general. Due to transcription conventions, suprasegmental features are affixed to segments in phonetic transcription and often used in the training procedure without criticism. This can lead to data sparsity when training triphone models. The ASR system inventories the number of phones/phonemes in the training data to determine the number of triphone models that should be estimated. Supposing that a given phone can have up to 5 separate diacritics affixed, this results in an upper bound of 120 different phone models for a single segment depending on the occurrences in the training data. In DanPASS, the fine-grained transcription results in a phone inventory of at least 400 distinct phones, which are different realisations of the segments. If the difference between two models is very small, the system will not be able to discriminate between the two realisations. Intuitively it seems an ASR systems would become more robust if a model trained for a segment encompassed all the different realisations of that segment rather than 120 models that describe the different realisations of one segment.

Transcription conventions for annotation of stress are also problematic in an ASR context. Stress is annotated with a diacritic that is affixed to a syllable, i.e. a series of segments. In IPA, the convention is to annotate stress as a prefix (ˈs i r k a) and in the Danish SAMPA alphabet, e.g. in DK-PAROLE, the convention is to annotate stress as a suffix (`s i r!  g a`). For an ASR system, this will be interpreted as another phone symbol and result in different phone sets.

Duration and syllabification are two features of interest. Syllabification in Danish usually occurs when a schwa is assimilated and a sonorant becomes syllabic. In Danish, syllabification most often occur as the result of schwa-assimilation (SA) and happens when the sonorant is not part of the onset (Schachtenhaufen, 2010). It is interesting to see what happens if the symbol is treated as a segment. It is easier to annotate and altering this transcription convention will decrease annotation time considerably. Schwa is the fourth most common segment in DanPASS and the combination of sonorant+schwa in the coda is quite common because of the Danish definite article suffix *-en* where assimilation is almost obligatory (Schachtenhaufen, 2010). An example of syllabification with SA can be seen in Table 1 as well as the dictionary entry used for the experiments.

No assimilation	l ɛ n ɡ d ə n
Syllabification (SA)	l ɛ n ɡ d n̩
Dictionary entry	`l E N d - n`

Table 1: Different phonetic transcriptions of the Danish word *længden* (the length).

Duration is of interest because, while it may or may not have prosodic information, it is affixed and related to a single segment. However, this annotation overlaps with the functionality of the transitions in Figure 1 which return to the same state and model the n-to-1 relationship between vectors and phone symbols. Since duration is implicitly modelled, it is interesting to see if filtering away duration can improve the WER and not introduce too much ambiguity. Omitting annotation for duration will also decrease annotation time.

5 Results

For this setup, CMU Sphinx (Placeway et al., 1997) is used to train an ASR system with SphinxIII as the decoder. SphinxIII is a C++ implementation and the main Large Vocabulary Continuous

[2] Note that not all diacritics are suprasegmental features, but all suprasegmental features are diacritics.

Speech Recognition (LVCSR) version. To ease reproducibility, the standard configuration of SphinxIII is used in all experiments together with a trigram language model trained on the orthographic material in DK-PAROLE. 3574 recordings or 80% of the corpus was used as training data and the remaining 20% as test set. The experiments were setup with a set of scripts from Henrichsen and Kirkedal (2011). Slight alterations were necessary to make the scripts handle IPA characters and the different tier names in the textgrids. The changes will be added to the github repository[3] when ready. DK-Parole has been used for preliminary studies due to the easy conversion from PRAAT textgrids to the sphinx training database layout, which did not entail any phonetic character conversion, only filtering. Also, the audio was recorded between 2006 and 2008 and the corpus contains the most recent speech recordings of standard Danish available at the time.

Evaluation Metrics:	WER	Words (3574)	Sentence Error	Sentences (894)
No diacritics	5.7%	691	34.3%	307
SA	5.6%	678	34.5%	308
Stød	5.9%	720	36.4%	325
+SA	5.7%	693	34.3%	307
Duration	5.9%	714	35.9%	321
+SA	5.8%	707	36.8%	329
Stress	6.0%	722	36.0%	322
+SA	5.9%	711	36.1%	323
Duration and stress	6.5%	791	38.0%	340
+SA	6.0%	734	36.7%	328
Stress and stød	6.5%	791	38.4%	343
+SA	6.3%	769	37.4%	334
Duration and stød	5.9%	719	36.4%	325
+SA	5.9%	720	36.4%	325
Duration, stress and stød	6.7%	808	38.0%	340
+SA	6.4%	775	38.1%	341

Table 2: Error rates for the DK-Parole ASR experiments. SA is short for schwa-assimmilation.

The WER and sentence error rates of 16 experiments were calculated using SClite (Fiscus, 1998) and are reported in Table 2. WER is the number of substitutions, deletions and insertions normalised by the number of words in the reference:

$$WER = \frac{Substitutions + deletions + insertions}{\# words} \tag{1}$$

Sentence error rate is also reported. In a post-editing scenario, it is relevant to know how many sentences needs to be edited. If a user of an ASR system needs to edit every sentence, the potential gains in efficiency will not be realised, but editing many mistakes in few sentences are not as time-consuming.

The experiments are grouped in pairs where the only difference in the paired experiments are the annotation of SA. In the sets of paired experiments, the transcription in the dictionary was

[3] `https://github.com/dresen/sphinx_setup`

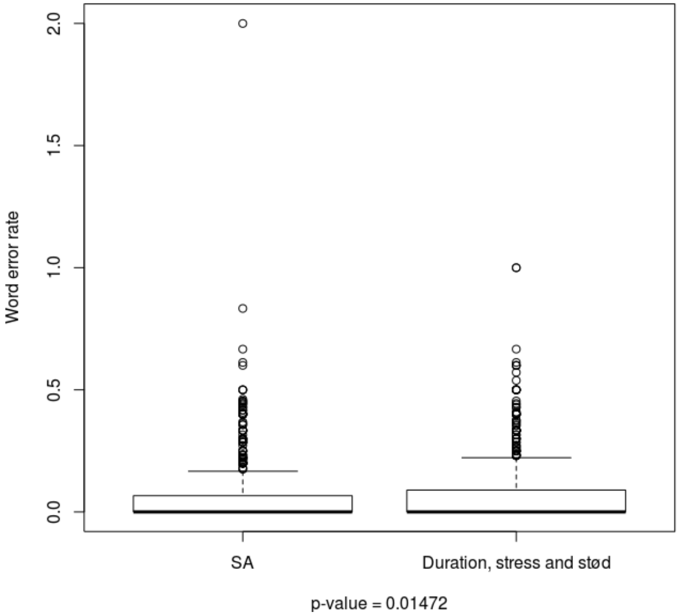

Figure 3: WER significance tests.

changed so all combinations of annotations for stress, duration, stød and SA have been carried out.

6 Discussion

Table 2 shows that including annotation for stress, stød, and duration increases the error rate. In each set of experiments, adding SA annotation has either no impact or decreases the reported error rates.

These preliminary results should be interpreted with caution. The training data is based on recordings of a single speaker reading text aloud. It is possible that this configuration of the pronunciation dictionary is only an improvement on this data set and does not generalise to other speakers.

Treating SA as a segment decreases the WER in the ASR experiments. This is so far the only case where the change from a diacritical to a segmental representation (from [n-] to [-] [n]) has improved the recognition rate. Adding a SA symbol and therefore an SA triphone model, seems to make the systems more robust. This is clear when comparing the WER of experiments including annotation for duration and stress with and without annotation for SA. Adding SA increases the recognition rate in seven out of nine comparisons.

Filtering out duration annotation improves the recognition rate by up to 0.5% WER. This indicates that the implicit modelling of time in triphone models are sufficient to model duration and the lack of this annotation does not introduce unresolvable ambiguity. The amount of ambiguity introduced and whether the ambiguity is resolved by the language model or elsewhere has not been determined. This is interesting because vowel duration is a distinctive lexical feature in Danish and all vowels in Danish exist in a long and short form (Grønnum, 2005). It also conflicts with the conclusions of Brøndsted and Madsen (1997) and indicates that long and

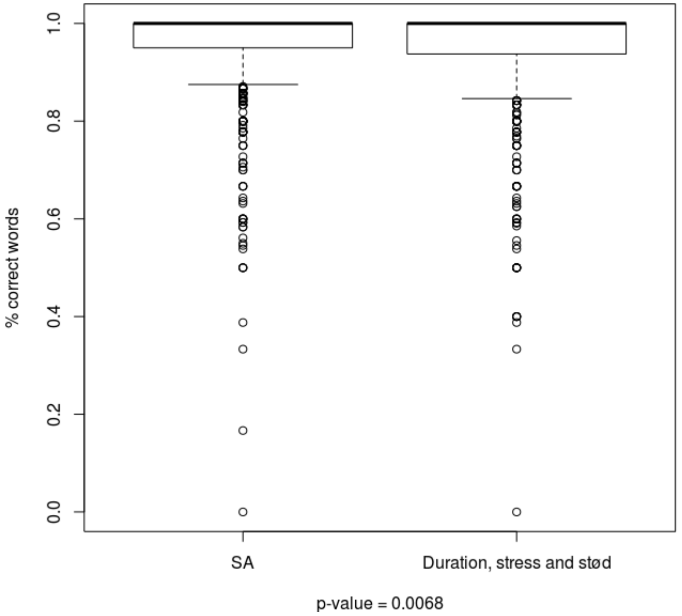

Figure 4: Number of correct words significance tests.

short versions of the same vowel symbol should be used to train a single model.

The changes in transcription give in small improvements. The combined improvement from the worst performing system to the best is statistically significant in terms of correctly recognised words and WER. The *p*-value and WER significance test can be seen in Figure 3.

A significance test for the number of correctly recognised words have been carried out as well. The results and *p*-value can be seen in Figure 4.

Some of the most difficult perceptual features to annotate are not salient for ASR systems. This conflicts with transcription conventions from the Danish phonetic community where these suprasegmental features are considered salient for human listeners. This could indicate that transcription can be handled by non-experts, e.g. via Amazon's Mechanical Turk to either create training data or acquire human evaluations (Novotney and Callison-Burch, 2010).

7 Conclusion and Outlook

Conventional speech transcription seems to be more fine-grained than necessary and annotation of suprasegmental features and duration in pronunciation dictionaries for state-of-the-art ASR systems did not improve the recognition rates. Currently, adding annotation for suprasegmental features are counter-productive. These features are important for human listeners, but the HMM and triphone structure of current ASR systems cannot take advantage of this information.

To verify these findings, it is important to repeat the experiments with a different corpus. With the very fine-grained transcription, multiple speakers and semi-spontaneous speech, DanPASS will make a good next step. The transcriptions in DanPASS will also make it possible to experiment with additional annotation.

References

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Brøndsted, T. and Madsen, J. (1997). Fonemteori og talegenkendelse. *Sprog og multimedier. Aalborg Universitetsforlag*.

Fiscus, J. (1998). Sclite scoring package version 1.5. *US National Institute of Standard Technology (NIST), URL http://www. itl. nist. gov/iaui/894.01/tools*.

Gregersen, F. (2007). The lanchart corpus of spoken danish, report from a corpus in progress. *Current Trends in Research on Spoken Language in the Nordic Countries*, 2:130–143.

Grønnum, N. (2005). Fonetik og fonologi, 3. udg. *Akademisk Forlag, København*.

Grønnum, N. (2006). Danpass-a danish phonetically annotated spontaneous speech corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genova, Italy, May*.

Henrichsen, P. (2007). The danish parole corpus-a merge of speech and writing. *Current Trends in Research on Spoken Language in the Nordic Countries*, 2:84–93.

Henrichsen, P. and Kirkedal, A. (2011). Founding a large-vocabulary speech recognizer for danish. In *Speech in Action*, pages 175–193.

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Ljolje, A. (1994). High accuracy phone recognition using context clustering and quasi-triphonic models. *Computer Speech & Language*, 8:129–151.

Novotney, S. and Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215. Association for Computational Linguistics.

Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., et al. (1997). The 1996 hub-4 sphinx-3 system. In *Proc. DARPA Speech recognition workshop*, pages 85–89. Citeseer.

Schachtenhaufen, R. (2010). Schwa-assimilation og stavelsesgrænser. *NyS*, (39):64–92.

Wells, J. et al. (1997). Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.