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.

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.

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.

```
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.*

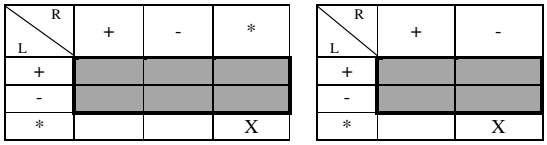
Group	Distinctive Features	Meaning
Tongue	High, Low, Front, Back [6]	Place of tongue in vowel
	Lateral, Anterior [6]	The tongue part and shape used to produce sound
	Dental [16] , Alveolar [16], Retroflex [19], Velar [16]	The tip/blade of tongue will be placed different places to form a constriction.
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.*

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



(a) Three-class setting (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)*

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.*

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 + $\Delta\Delta$). 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.

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)

$$D_i = \frac{1}{N}\sum_{j=1}^{N}\left(T_j - A_{i,j}\right) \quad (1)$$

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

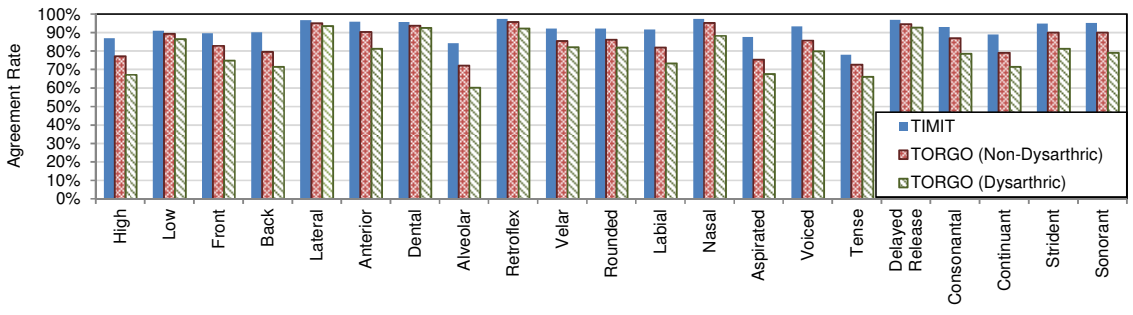


Figure 4: *The agreement rate of each DF between recognized results and canonical DFs.*

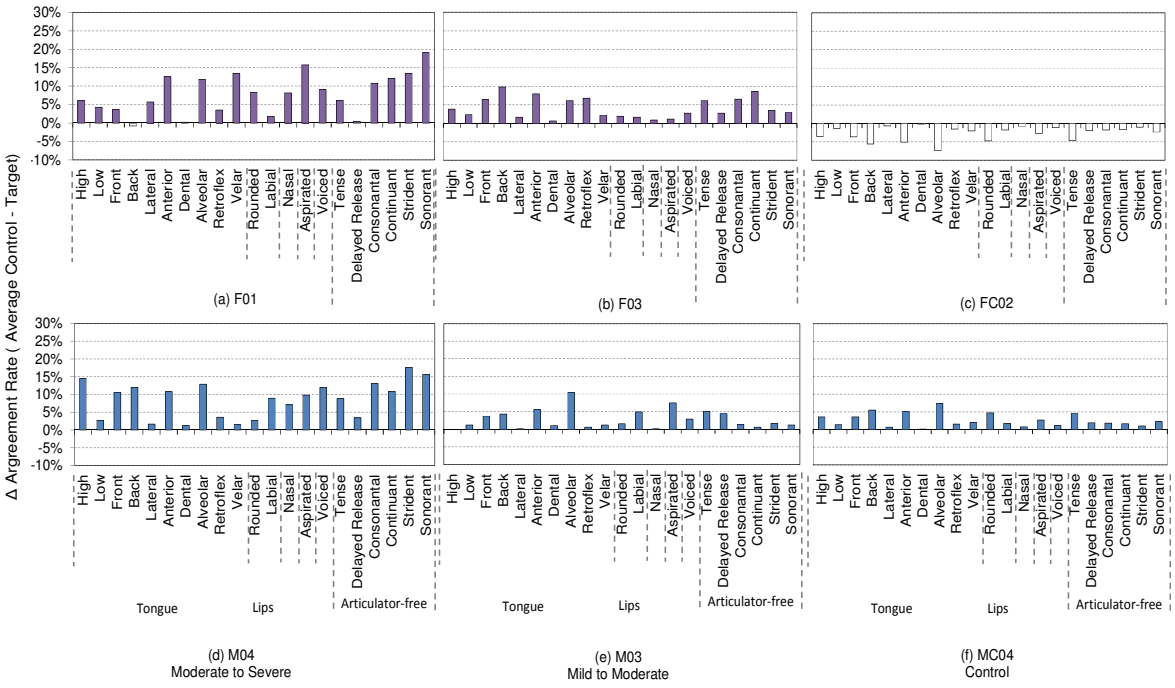


Figure 5: *The difference between the average DF agreement rate from the control subjects and the corresponding DF agreement rate of each dysarthric subject. The agreement rates of most of the DF drop substainally for severely dysarthric subjects. The agreement rates in moderately and mildly dysarthric 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 com-

plete 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.