EXPLORING ARTICULATORY CHARACTERISTICS OF CANTONESE DYSARTHRIC SPEECH USING DISTINCTIVE FEATURES

Ka Ho Wong [1], Wing Sum Yeung [1], Yu Ting Yeung [2] and Helen Meng [1,2]
[1] Human–Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
[2] Stanley Ho Big Data Decision Analytics Research Centre,
The Chinese University of Hong Kong, Hong Kong SAR, China
khwong@se.cuhk.edu.hk, wsyeung@se.cuhk.edu.hk, ytyeung@se.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk

ABSTRACT

Dysarthria is a kind of motor speech disorder due to neurological deficits. Understanding the articulatory problems of dysarthric speakers may help to design suitable intervention strategies to improve their speech intelligibility. We have developed an automatic articulatory characteristics analysis framework based on a distinctive feature (DF) recognition. We recruited 16 Cantonese dysarthric subjects with spinocerebellar ataxia (SCA) or cerebral palsy (CP) to support our research. To the best of our knowledge, this is among the first efforts in collecting and automatically analyzing Cantonese dysarthric speech. The framework shows a close Pearson correlation to manual annotation of the subjects in most DFs and also in the average DF error rates. It indicates a potential way to describe articulatory characteristics of dysarthric speech and automatically assess it.

Index Terms— dysarthria, Cantonese, speech disorder, distinctive features, multilayer perceptron

1. INTRODUCTION

Dysarthria is a kind of motor speech disorder due to neurological deficits. The speech disorder affects the communication of the daily lives of the patients. Problems of dysarthric speech include but are not limited to imprecise consonants, distorted vowels and prosody problems such as monopitch [1]. A detailed analysis of the articulatory characteristics of dysarthric speech potentially facilitates speech therapists to design suitable intervention strategies.

The analysis of dysarthric speech covers articulation, prosody, speaking style and other aspects. In [2], the authors studied the speaking rate and style of French dysarthric speech recorded from subjects with Parkinson's disease, amyotrophic lateral sclerosis (ALS) and cerebellar ataxia subjects. The phonetic analysis of English dysarthric speech associated with CP and ALS was performed in [3]. For Cantonese, the dysarthria associated with CP was examined manually in [4] on the aspects of phones, tones, places and manners based on 22 subjects.

The characteristics of dysarthric speech may depend on the region of brain lesion, which in term depends on the etiology. The affected brain region depends on the type of disease. For subjects with ALS, the upper and lower motor neurons are under progressive degeneration [1]. For subjects with SCA, there is degeneration in the cerebella. For CP subjects, several brain regions are affected.

Our study focuses on articulatory characteristics of dysarthric speech. By representing articulatory features with DFs, our approach is to develop automatic DF detectors. Previously, we have explored this framework based on the English corpus [5] and have obtained reasonable results. In this paper, we investigate the applicability of our framework on Cantonese dysarthric speech.

This paper is organized as follows: A brief description of Cantonese and distinctive features are in Section 2 and Section 3 respectively. Section 4 presents the collected speech and annotation. We present the work on automatic detection of DFs in Section 5. Section 6 presents the evaluation of our framework. Finally, we conclude and present our future directions in Section 7.

2. CANTONESE PHONOLOGY AND PHONETICS

Each Cantonese character is pronounced as a single syllable with a lexical tone [6]. A base syllable refers to the tone-independent syllable. A base syllable is divided into two parts: the initial and the final. The initial is an optional consonant called the onset. The final consists of a vowel nucleus and a consonant coda. The onset and the coda are optional. Tone is a supra-segmental component that is also a syllable characteristic. In this work, we focus on the base syllable. For labelling the Cantonese syllables, we adopt the phonetic symbols in the Jyutping system developed by the Linguistic Society of Hong Kong [7]. The phonetic symbols in Jyutping system will be used in this paper.

3. DISTINCTIVE FEATURES

Distinctive features (DFs) include articulator-bound and articulator-free features, and aim to represent the articulatory corresponding to different acoustic patterns [8]. Table 1 lists the 21 DFs used in this work. Each stop is split into a closure and a release. Each diphthong is split as two vowels [5]. Each phone, including the split phones, is mapped into a DF vector by table lookup. Each DF has four possible values: positive ("+"), negative ("–"), irrelevant ("x") and

Group	DFs	Brief Meaning
Tongue	Coronal	Tongue blade is raised towards the teeth or the hard palate
	High, Low, Front, Back	Position of the tongue
	Lateral	How the tongue manipulates the airstream flow
	Tense	Tongue configuration with a greater constriction
	Velar [15], Alveolar [15]	Place of obstruction made by the tongue
Lips	Labial	Constriction at the lips
	Rounded	Protrusion of the lips
Tongue / Lips	Anterior	Horizontal position of the primary constriction
Soft Palate	Nasal	Soft palate is lowered
Vocal cords	Spread glottis	Vocal cords are drawn apart
	Voiced	Vocal cords vibrate periodically
Articulator−free	Syllabic	Constitution of syllable peaks
	Consonantal	Sustained vocal tract constriction
	Sonorant	Vocal tract configuration is open
	Continuant	Vocal tract configuration allows the airstream to flow through the center of the oral tract
	Strident	A constriction forces the airstream to strike two surfaces
	Delayed Release [16]	Vocal tract closure released with a delay

Table 1: *Definitions of the 21 distinctive features (DFs) used in this work, following* [9].

unspecified ("/"). For examples, an "unspecified" DF is [NASAL] for /h/ [9] — where nasalization has no effect on the identification or recognition of /h/. An example of an "irrelevant" DF is [TENSE] for /p/ — [TENSE] describes a greater degree of constriction with a tongue body or tongue root and does not play a part in the articulation of /p/. In this work, we focus only on DF values that are either positive or negative.

4. CORPUS DESCRIPTION

4.1. Prompt and Subjects

In order to prepare for the collection of Cantonese dysarthric speech, we have designed a set of recording prompts that cover a range of speaking styles, including single words, short sentences, paragraphs, articulatory tasks and conversations. We recruited 13 SCA subjects, 3 CP subjects for dysarthric speech recordings and 9 non-dysarthric subjects for recording. The current investigation takes an initial step to look at the articulatory characteristics of Cantonese dysarthric speech. Hence we focus on the single words (61 totals) which fully cover the syllable initials and finals in Cantonese. The non-dysarthric speech in our collection includes 366 training utterances (3 males, 3 females) and 183 testing utterances (1 male, 2 females) without any overlapped subjects [10].

	O	N	O	N	C	O	N
Canonical			h	a	k	j	e
Annotation 1 (a1)			t	aa	n	j	e
Annotation 2 (a2)	t	oi	w	aa	k	j	e

Figure 1: *Two annotations from two annotators for the canonical pronunciation /hak/ /je/ (黑夜). O, N and C are abbreviations for onset, nucleus and coda. In both annotations, /a/ is aligned with /aa/ and the two phones differ in terms of the feature [TENSE]. In addition, /h/ is aligned with /t/ in the first annotation (a1). In Annotation 2, the canonical /h/ is aligned with the onset /w/ because they share more commonalities in articulation − both /h/ and /w/ are [+CONTINUANT].*

4.2. Manual Annotation of Cantonese Phones

Undergraduate students studying in the Department of Linguistics and Modern Languages, and the Department of Chinese Language and Literature are recruited as annotators. The students are familiar with the Cantonese pronunciation system. However, they have no previous experience in labeling dysarthric speech. The annotators listen to the utterances with a Sennheiser PC155 headset, and annotate the utterances with Jyutping syllables **[7]**. We do not provide information about text prompts of utterances, and we do not limit the maximum number of times that they can listen to the utterances.

A trained linguist aligned the human phonetic annotations to canonical pronunciation manually. We prepared the canonical pronunciation of each utterance according to the Cantonese pronunciation dictionary **[11]**. The canonical nucleus is aligned with the annotated nucleus first. Then the canonical onset and coda are aligned with the annotated onset and coda respectively. If there are multiple choices of annotated onsets, nucleus or coda, the one with minimum number of DF value differences (changing from "+" to "−" and vice versa, i.e. [+|−] → [−|+], or [+|−] → [x] or [x|/] → [/|x]) will be chosen. An example is shown in Figure 1.

4.3. Substitutions and Deletions DF

Next we compare the DFs of each canonical phone with the DFs of the aligned, manually labeled phone. If the DF values differ ([+|−] → [−|+], or [+|−] → [x]), then we consider that a DF substitution error has occurred. If no labeled phone is aligned to a canonical phone, then the DFs of the canonical phone are considered as deletion errors. The DF error rate $S_{j,m}$ of each DF value j for subject m is defined as:

$$S_{j,m} = \frac{E_{j,m}}{N_{j,m}} \quad (1)$$

where $E_{j,m}$ is the number of substitution or deletion errors for DF value j in subject m, $N_{j,m}$ is the total number of phone segments including the DF value j produced by subject m.

As an example, consider Figure 1. The canonical nucleus is /a/ which is [−TENSE]. However, both annotators labeled it as /aa/ which is [+TENSE]. This suggests that the articulation of [−TENSE] may be problematic for this speaker, and the error rate of [−TENSE] is 2/2 (i.e., 100%). As another example, consider the canonical phone /h/ in Figure 1,

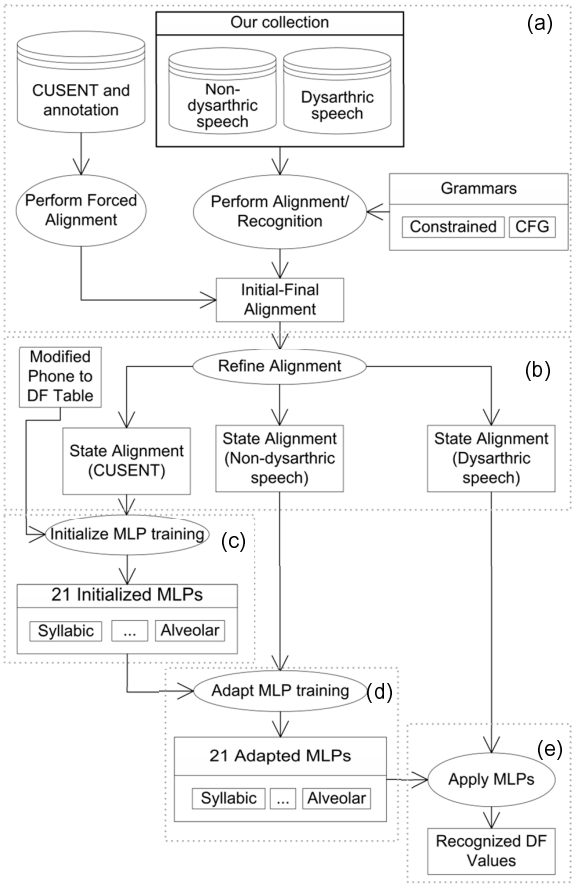

Figure 2: *The flow diagram for DF recognition.*

which is [+CONTINUANT]. In the first annotation, /h/ is aligned with /w/ which is also [+CONTINUANT]. In the second annotation, /h/ is aligned with /t/ which is [−CONTINUANT]. Hence the substitution for this subject for [+ CONTINUANT] is counted as 1/2 and is considered less serious compared with the previous [−TENSE] feature.

5. AUTOMATED ANALYSIS OF DYSARTHRIC SPEECH BASED ON DISTINCTIVE FEATURES

Figure 2 shows the flow diagram of automated analysis of dysarthric speech. First, the acoustic signal is aligned with the canonical pronunciation at the phone-level (initials and finals) as in Figure 2 part a. Second, the phones in the alignment are modified and mapped into DF values by table lookup (Figure 2 part b). An initial MLP is trained on CUSENT [12], a Cantonese non-dysarthric speech corpus (Figure 2 part c) and adapted with the collected non-dysarthric speech (Figure 2 part d). The adapted MLP model is applied to dysarthric speech for DF recognition (Figure 2 part e). Finally, recognized DF values are compared with canonical DF values. The detail of each step is discussed below.

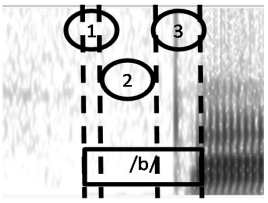

Figure 3: *The 3 states of stop /b/ from the non-dysarthric subject C002F. The 1 state will be the closure, the 2 state is a transition and the 3 state is the release part.*

5.1. Phone Alignment

Alignment of canonical phones with dysarthric speech presents several challenges due to characteristics such as phone insertion, slow speaking rates, prolonged pause intervals, disfluencies, stuttering and pronunciation deviations. For example, we have a subject (coded S0014M) who deleted the phone /g/ in the syllable /gang/ and pronounced /ang/. In order to adapt our alignment system for dysarthric speech, we added a constrained grammar which allows optional phone deletions to perform automatic forced alignment. Details can be found in [5] based on our work in English.

Forced alignment is performed with the HTK toolkit [13] with a well−trained acoustic model (AM) [12]. The AM is trained with the CUSENT Cantonese speech corpus according to Jyutping phonetic transcription. The AM follows hidden Markov model (HMM) topology and with 32 Gaussian mixtures per state.

To further align the stops, diphthongs and finals, the state−level alignment is applied. For example in Figure 3, state 1 of /b/ is aligned to the closure part of /b/. State 2 is considered a transition from closure to release. State 3 is considered as the release of /b/. The aligned phone will be mapped into DF (Figure 2 part b) by table lookup.

5.2. Multilayer Perceptron for DF Recognition

To train a DF recognition system, we start from a non-dysarthric speech corpus, CUSENT [12] as in Figure 2 part c. CUSENT is a read speech corpus of continuous Cantonese sentences. It includes 20,400 training utterances from 68 subjects and 1,200 test utterances. The corpus contains phonetic-level annotations.

An initial frame-based MLP classifier for each DF is trained with the CUSENT training data. 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 [5]. The input layer that takes in the 39-dimenensional Mel-frequency cepstral coefficients (MFCC) (12 coefficients + log-energy + Δ + ΔΔ) with 351 input nodes, three hidden layers with 500 x 120 x 500 nodes and an output layer with 2 nodes to capture both "+" and "−" values of a DF. Only a frame lying in the central of a phone is interested. During the training of each DF, we skip the frames which are silent or labeled as unspecific or irrelevant, but we still include them in the input vectors.

In Figure 2 part d, we further adapt the MLP classifiers with non-dysarthric speech data of our collection. The initial weights of the adapted classifiers are the same as the weights of the initialized MLP classifiers. The weights are updated with the same training process.

5.3. DF Recognition by MLP

We analyzed the outputs of the DF based on phones in CUSENT with "irrelevant". In these situations, we found that the output nodes generally produce low values for both the "+" and "−" output nodes. Hence, we devise our approach such that when the output nodes both have values lower than a threshold, the detected DF value is classified as "irrelevant". To define the threshold, we normalize the values of the output nodes of DF j with a standard score (z−score), z_j, according to the following equation:

$$z_j = \frac{x_j - \bar{x}_j}{\sigma_j} \quad (2)$$

where x_j is the value of "+" or "−" node of DF j, $\bar{x}_j$ and σ_j are the corresponding mean and standard deviation of the positive/negative node of DF j.

The threshold is set to be −1 empirically. For frames where the z-score of both output nodes lie above the threshold, the node with the higher z-score is selected as the final output for that DF.

6. DISCUSSIONS

6.1. DF Matchings of Individual Phone

A DF mismatch is defined if any annotator perceived mismatch in DF produced. F1 scores among different DFs ranged from 0.069 to 0.401, with an average of 0.268. Distorted phone productions may lead to various perceptions.

6.2. DF Recognition Performance of Individual Subject

We have explored the relationships among DF errors for each subject. We calculate the substitution error rate by equation (1) based on the manual annotation and MLP results. The Pearson correlations of each DF are shown in Figure 4. The average of correlation is 0.7.

The correlation of most DFs have 0.6 or above. It shows that most DFs can capture the same error trend comparing with manual annotation. The lowest three DFs are [DELAYED RELEASE], [LOW] and [STRIDENT]. [DELAYED RELEASE] is related to timing of vocal tract closure release. The prolonged pronunciation may affect the DF recognition. The number of "+" and "−" samples for [LOW] and [STRIDENT] is unbalanced. The number of [−LOW] samples is much more than [+LOW]. It may lead the MLP to bias towards one label. The error rates of [STRIDENT] based on MLP results and manual annotation have a large difference for four subjects (coded S0007M, S0014M, S0015M and S0017M). A large number of deleted [+STRIDENT] based on manual annotation occurred in their recording. The deletion affects the forced alignment accuracy, which in turn affects DF classification accuracy.

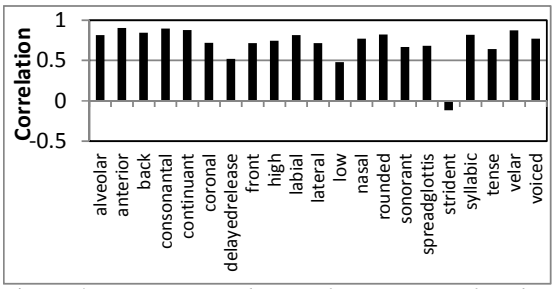

Figure 4: *Pearson correlation of error rates of each DF from MLP and manual annotation for dysarthric subjects.*

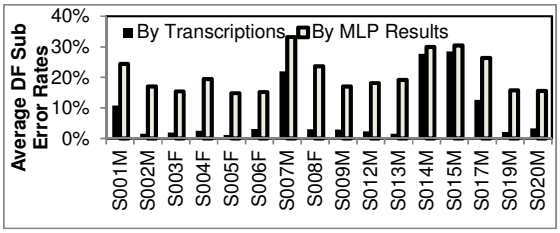

Figure 5: *The average DF substitution error rate comparison between automatic analysis and annotations.*

6.3. Average DF Error Rates of Individual Subject

We further analyze the average DF substitution error rates based on the automatic analysis and annotations for each subject (Figure 5). The Pearson correlation is 0.9. The automatically generated average DF error rates are higher than the error rates based on annotation. However, there is good correlation between the manually labeled and automatically classified DF. We believe that we can improve the DF classification performance by improving the automatic alignment between the canonical phones and the acoustic signal to get more accurate phone boundaries.

7. CONCLUSIONS AND FUTURE WORK

In this work, we have developed an automatic framework of analyzing the articulatory characteristics for dysarthric speech based on distinctive features. Results show that it provides a highly correlated result compared to manual annotations on most DFs. It is also possible to assess the dysarthric speakers in terms of average DF substitution error rates. The framework can help describe the articulatory problems of dysarthric speech. In the future, we will try to improve MLP accuracy by applying the forced alignment described in [14].

8. ACKNOWLEDGEMENTS

This project is partially sponsored by a grant from the Hong Kong SAR Government General Research Fund (reference no. GRF415513).

9. REFERENCES

[1] D. B. Freed, Motor Speech Disorders: Diagnosis & Treatment, Clifton Park: Delmar, Cengage Learning, 2012.

[2] B. Bigi, K. Klessa, L. Georgeton and C. Meunier, "A Syllable-Based Analysis of Speech Temporal Organization: A Comparison Between Speaking Styles in Dysarthric and Healthy Populations," in *Interspeech*, Dresden, Germany, 2015.

[3] K. Mengistu and F. Rudzicz, "Adapting Acoustic and Lexical Models to Dysarthric Speech," in *IEEE International Conference of Acoustic, Speech and Signal Processing*, 2011.

[4] T. L. Whitehill and V. Ciocca, "Speech Errors in Cantonese Speaking Adults with Cerebral Palsy," *The Clinical Linguistics and Phonetics,* vol. 14, pp. 111-130, 2000.

[5] K. H. Wong, Y. T. Yeung, P. C. M. Wong, G.-A. Levow and H. Meng, "Analysis of Dysarthric Speech Using Distinctive Feature Recognition," in *the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Dresden, Germany, 2015.

[6] P. C. Ching, T. Lee, W. K. Lo and H. Meng, "Cantonese Speech Recognition and Synthesis," in *Advances in Chinese Spoken Langauge Processing (C.H. Lee, H. Li, L.S. Lee and R.H. Wang, Eds.)*, World Scientific Publishing Company, 2007, pp. 365-386.

[7] The Linguistic Society of Hong Kong, "The Jyutping Scheme," 1993. [Online]. Available: http://www.lshk.org/node/47. [Accessed 11 December 2014].

[8] K. N. Stevens, Acoustic Phonetic, Cambridge: MIT Press, 1998.

[9] M. Halle and G. N. Clements, Problem Book in Phonology: A Workbook for Introductory Courses in Linguistics and in Modern Phonology, Cambridge: MIT Press, 1983.

[10] K. H. Wong, Y. T. Yeung, E. H. Y. Chan, P. C. M. Wong, G.-A. Levow and H. Meng, "Development of a Cantonese Dysarthric Speech Corpus," in *Interspeech*, Dresden, Germany, 2015.

[11] "Chinese Character Database: With Word-formations Phologoically Disambiguated According to the Cantonese Dialect," Research Centre for the Humanities Computing, The Chinese University of Hong Kong, Januaru 2003. [Online]. Available: http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/. [Accessed June 2015].

[12] T. Lee, W. K. Lo, P. C. Ching and H. Meng, "Spoken Language Resources for Cantonese Speech Processing," *Speech Communication,* vol. 36, no. 3-4, pp. 327-342, 2002.

[13] S. Young, J. Odell, D. Ollason, V. Valthcey and P. Woodland, The HTK Book, Cambridge University, 1995.

[14] Y. T. Yeung, K. H. Wong and H. Meng, "Improving Automatic Forced Alignment for Dysarthric Speech Transcription," in *Interspeech*, Germany, 2015.

[15] P. Ladefoged and K. Johnson, A Course in Phonetics, Boston: Wadsworth, Cengage Learning, 2009.

[16] N. Chomsky and M. Halle, The Sound Pattern of English, Harper & Row, 1968.