Development of a Cantonese Dysarthric Speech Corpus

Ka Ho Wong[1], *Yu Ting Yeung*[2], *Edwin H. Y. Chan*[3], *Patrick C. M. Wong*[4], *Gina-Anne Levow*[5] *and Helen Meng*[1,2]

[1] Human-Computer Communications Laboratory,
[1]Department of Systems Engineering and Engineering Management
[2] Stanley Ho Big Data Decision Analytics Research Centre,
[3] School of Life Sciences,
[4]CUHK-Utrecht University Centre for Language, Mind and Brain,
[4]Department of Linguistics and Modern Languages,
[1,2,3,4]The Chinese University of Hong Kong, Hong Kong SAR, China
[5]Department of Linguistics, University of Washington, Seattle, WA USA

khwong@se.cuhk.edu.hk, ytyeung@se.cuhk.edu.hk, hyechan@cuhk.edu.hk, p.wong@cuhk.edu.hk, levow@uw.edu, hmmeng@se.cuhk.edu.hk

Abstract

Dysarthria is a neurogenic communication disorder affecting speech production. Significant differences in phonemic inventories and phonological patterns across the world's languages render generalization of disordered speech patterns from one language (e.g, English) to another (e.g., Cantonese) difficult. Capitalizing on existing methods in developing English-language dysarthric speech corpora, we develop a Cantonese corpus in order to investigate articulatory and prosodic characteristics of Cantonese dysarthric speech, focusing on speaking rate and pitch and loudness control. Currently, we have collected 7.5 and 2.5 hours of speech data from 11 dysarthric subjects and 5 control speakers respectively. Our preliminary analysis reveals the characteristics of Cantonese dysarthric speech are consistent with general properties of motor speech disorders found in other languages.

Index Terms: Dysarthria, motor speech disorders, Frenchay Dysarthria Assessment, regression, corpus, Cantonese

1. Introduction

Dysarthria is a motor speech disorder caused by neurologic deficits such as cerebral palsy (CP), stroke, and neurogenerative diseases such as spino-cerebellar ataxia (SCA). In Hong Kong, Yam *et. al.* shows that 1.3 per 1,000 children aged 6 to 12 years has cerebral palsy [1]. Chau *et. al.* also shows a continuous increase in hemorrhagic stroke incidence among adults aged from 35 to 44 during 1999 to 2007 [2]. Dysarthria severely affects daily life. Although speech technologies such as speech recognition and speech re-construction may help improve the quality of life of the subjects, the prerequisite of these technologies is corpus data which can assist in capturing characteristics of dysarthric speech in different languages with different articulatory features. While several studies have investigated dysarthric speech in English, corpus data for other languages are generally lacking.

Here, we provide a preliminary report on our effort in developing a corpus of dysarthric speech in Cantonese. Significant differences in phonological structures between English and Cantonese including phonemic inventories, phonotactics, tonal patterns and syllable structure [3], allow us to examine language-specific and language-general characteristics of dysarthric speech in order to facilitate the development of speech applications to support clinical assessment and rehabilitation.

The collection of English dysarthric speech has been in progress for close to two decades [4] [5] [6]. However, the currently available English dysarthric speech corpora are unsuitable for dysarthric speech research for other languages, due to different articulation features. For Cantonese, the lack of publicly available dysarthric speech corpora motivates us to collect one to support our research, in areas such as dysarthric speech error analysis and assessment.

In the next section, we will review the available dysarthric speech corpora in English and other languages. In Section 3, we will discuss the design and the collection process of our new Cantonese dysarthric speech corpus. In Section 4, we will describe our setting of automatic forced alignment for phone-level time alignments of speech data. We will also discuss some preliminary analysis in Section 5. Finally, we will conclude and present our future directions in Section 6.

2. Available dysarthric speech corpora

Three English dysarthric speech corpora are publicly available. The first is the Nemours corpus [4], which aims at testing the intelligibility of dysarthric speech before and after enhancement by various signal processing methods. The corpus includes 11 male dysarthric subjects. The results of the Frenchay Dysarthria Assessment version 1 (FDA-1) [7], one of the standardized assessments of English dysarthric speech, are reported in the corpus. The stimuli of the Nemours corpus include non-sense sentences (e.g. "The sin is sitting the who"), and frequently used passages for phonetic studies (e.g. the rainbow passage and the grandfather passage) [8].

The second is the Universal Access-Speech (UA-Speech) corpus [5]. It aims to promote the development of user interfaces for dysarthric subjects. The corpus includes 13 male and 4 female dysarthric subjects and 9 male and 4 female non-dysarthric subjects. The overall speech intelligibility of each subject is included. The speech intelligibility is obtained from how many words can be recognized correctly by human without prior knowledge of prompt. UA-Speech corpus includes digits, computer commands (e.g. delete), radio alphabet letters (e.g. “alpha” and “bravo”), common (e.g. “it” and “you”) and uncommon words (e.g. “moonshine” and “naturalization”).

The TORGO corpus aims at the development of automatic speech recognition using acoustic and articulatory features [6]. The corpus includes 5 male and 3 female dysarthric subjects and 4 male and 3 female non-dysarthric controls. FDA-1 results of dysarthric subjects are included. TORGO includes non-words (e.g. “ah-p-eee”), short words and restricted sentences.

The project Quality-of-Life technology (QoLT) focuses on developing a Korean speech recognition system for dysarthric speech [9]. The project includes 65 male and 35 female dysarthric subjects and 20 male and 10 female non-dysarthric controls. The corpus also includes speech intelligibility scores of each subject. The prompts include words, Korean phonetic alphabets and code words.

For Cantonese, the work of [10] aims at characterizing common speech errors using a perceptual-phonetic approach at single word level. The researchers collected speech data of 12 male and 10 female Cantonese dysarthric subjects with cerebral palsy. The prompts include 100 monosyllable words such as 樹 (/syu6/, “tree”), and 瓜 (/kwa1/, “melon”). Each initial, final, vowel and tone in Cantonese appeared a minimum of three times. However, the corpus is not publicly accessible. In additional to monosyllablic words, we also want to investigate the speech production performance in different contexts.

3. Design of Cantonese Dysarthric Speech Corpus

3.1. Task Design

The stimuli we selected are structured to include a range of speaking styles like single word, short sentence, paragraph and conversation, as well as articulatory tasks. The articulatory tasks can help to isolate specific articulatory or prosodic performance characteristics such as the pitch control ability. The fifteen tasks performed by our subjects are as follows:

Words	Syllables (Jyutping)	Translation
出口	/ceot7 hau2/	exit
黑夜	/hat7 je6/	dark night
春天	/ceon1 tin1/	spring (season)
打風	/daa2 fung1/	typhoon approaching
合桃	/hap9 tou4/	walnut
間尺	/gaan3 cek2/	ruler
生活	/sang1 wut9/	living
脫落	/tyut8 lok9/	fall off
客廳	/haak8 teng1/	living room
髮夾	/faat8 gep2/	hair pin
喵	/meu1/	cat meow
舐舐脷	/lem2 lem2 lei6/	lick the lips

Table 1: *Examples of word-level stimuli.*

Sentences (Syllables in Jyutping)
亞公食飯。/aa3/ /gung1/ /sik6/ /faan6/ (“Grandfather, enjoy meal.”)
姑姐喺邊？ /gu1/ /ze1/ /hai2/ /bin1/ (“Aunt is where?”)
件衫愈著愈闊。/gin6/ /saam1/ /jyut6/ /zoek3/ /jyut6/ /fut3/ (“The clothes get wider after dressing.”)

Table 2: *Examples of short sentences in part 2 and their translations.*

- Task 1: Word-level stimuli. Task 1 is adopted from the Hong Kong Cantonese Articulation Tests (HKCAT) which includes 49 words [11]. 12 extra words as shown in Table 1 are added to the task to cover all Cantonese initials and finals in the Jyutping system [12]. Task 1 consists of a total of 61 prompts.
- Task 2: Sentence-level stimuli. The prompts of Task 2 consist of short sentences (Table 2). A set of 23 short sentences are designed where the number of syllables of each sentence is similar to the sentence test in Frenchary Dysarthria Assessment (FDA-2) [13] and the sentences cover all Cantonese initials, finals and lexical tones. The average number of characters per sentence is 4.56 (standard deviation: 0.82).
- Task 3: Paragraph-level stimuli. The prompt of Task 3 is a phonetically rich passage widely used by Cantonese speech therapists[1]. The passage contains 121 Chinese characters.

The prompts of Task 4 to 14 are motivated by the FDA-2. FDA-2 includes 7 sections, namely *reflexes, respiration, lips, palate, laryngeal, tongue* and *intelligibility*. Each section includes several tasks. Some tasks in FDA-2 are non-speech tasks, such as assessment of coughing difficulty. These tasks are not included in our data collection. Task 4 corresponds to the respiration section of FDA-2. Task 5-14 correspond to the lips, palate, laryngeal and tongue sections of FDA-2. The Cantonese stimuli are designed to have similar phonetic characteristics. For example, we follow the similar occurrence frequency of bilabial or labiodental phonemes as in the sentences of FDA-2. Table shows the examples of stimuli from Tasks 4 to 14.

Task 15 is a 5-minute conversation. The operator chats with subjects about casual daily topics. The responses of the subjects are recorded.

3.2. Data Collection Setting

Recording takes place in a quiet meeting room. The samples are recorded with a Shure headset SM10A microphone, connected to a notebook through a Roland Quad Capture interface. The distance of the microphone from the mouth corner is around 1.5 cm. The stereo audio is sampled at 44.1 kHz and quantized at 16 bits.

A monitor is placed in front of the subject. A digital camera, Panasonic G6X, is put on the top of monitor and focuses on the face of the subject. Since the lip area is very small and the lips may shift out of view when the subject moves their head slightly, the digital camera captures the whole face rather than just the lip movement (Figure 1). Video data are only recorded with the subject’s consent. The video quality is high-definition (1920 x 1080 pixels) with 50 frames per second (FPS). We use high FPS so the fast movement of the lips can be captured. All window curtains are closed to ensure consistent lighting. The Roland Quad Capture interface directs the audio input from the Shure headset to both the notebook and the digital camera. The audio input is thus

[1] Chinese text available at: http://www1.se.cuhk.edu.hk/~khwong/index_files/Page327.htm

Task	Reference	Task Examples
4	Respiration	Please count from one to twenty in one breath as fast as possible.
5-7	Lips	Please say /baa1/ ten times as fast as possible.
8-9	Palate	Please take a breath through your nose and say /aa1/ /aa1/ /aa1/. Repeat four times. Please say the following pair of words • 內 /noi6/ (inside) 代 /doi6/ (generation) • 扭 /nau2/ (twist) 斗 /dau2/ (water dipper)
10-12	Laryngeal	Please say “Do Re Mi Fa Sol La” from the lowest tone to highest tone stepwise. Please say one two three four five six from the lowest loudness to highest loudness stepwise.
13-14	Tongue	Please say 「大哥彈琴得 ，打關斛都得。」(Brother plays piano well, also do somersault well) /daai6/ /go1/ /taan4/ /kam4/ /dak1/ /daa2/ /gwaan1/ /dau2/ /dou1/ /dak1/

Table 3: *Examples of speech production tasks.*

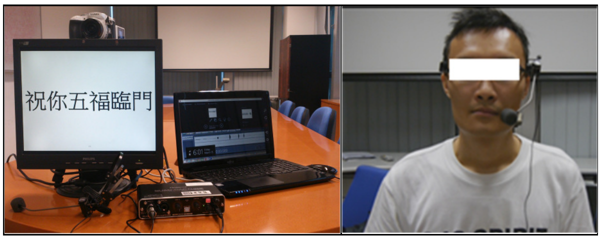

Figure 1: *The recording setup (Left). The subject sits in front of a large screen. The sample video captured from the digital camera is shown on right.*

synchronized between the notebook and the digital camera. Figure 1 illustrates the whole setup.

3.3. Data Collection Process

Subjects read aloud the prompts displayed on the screen. Each time, the screen displays only one prompt. For passages which are too long fit on to a single screen, the prompts are spread over several pages. The font size is maximized on the screen as dysarthric subjects may also have visual impairments. Each prompt is recorded at least twice. If the prompt is misread (e.g. missing characters), the subject reads aloud the prompt again to ensure that the mistake is not due to carelessness. For the task which requires the subjects to sing different tones, we allow them to try several times to ensure that they make their best effort to produce the highest tone. For Task 4 to 14, reference videos are shown to the subjects if necessary. For the 5-minute conversation in Task 15, we try to raise open-ended questions and let the subjects elaborate. Breaks and water are provided to prevent vocal fatigue and to maintain consistent speech quality across the tasks.

Slow speaking rate of dysarthric patients makes longer time to complete all tasks. Control subjects need around 30 to 40 minutes to complete all tasks. Dysarthric patients need around 1 hour to 1.5 hours. In the reading passage task 3, some dysarthric patients will miss some characters even tried two times.

3.4. Subjects

Our target corpus size is 100 dysarthric patients and at least 30 control subjects. The distribution of the subjects and controls recorded to date is shown in Table 4. All are from Hong Kong and native speakers of Cantonese. All subjects completed all tasks. The dysarthric subjects are diagnosed with cerebellar degeneration. Ten are due to spino-cerebellar ataxia (SCA) and one is due to another disease. Injury to the cerebellum can result in ataxic dysarthria [14]. 7.5 hours of dysarthric speech and 2.5 hours of control speech have been collected so far as shown in Table.

Subjects	Gender	Amount	Age	Total
Control	Male	3 (1.6 hrs)	25 – 30	5
	Female	2 (0.9 hrs)	26, 38	
Dysarthric	Male	6 (3.8 hrs)	33 - 63	11
	Female	5 (3.7 hrs)	46 - 67	

Table 4: *The distribution of 16 subjects.*

4. Forced Alignment of Speech Data

Manual alignment of the phone boundaries is a time consuming process. We apply automatic forced alignment with the HTK toolkit [15] to obtain the time-aligned phone-level transcriptions of the collected audio data. At the current stage, the corpus only covers a few hours of speech and a few speakers. This amount of data is insufficient for training a Cantonese acoustic model for automatic alignment. We develop speaker-independent acoustic models from CUSENT [16], a Cantonese read speech corpus for automatic forced alignment. The training set of the CUSENT corpus consists of about 20 hours of read speech data from 68 speakers. We have developed a monophone acoustic model according to the Jyutping pronunciation scheme. The acoustic model is based on hidden Markov model (HMM) architecture, with 128-component Gaussian mixture model as emission probability distribution at each state.

5. Preliminary Analysis

Prior work [14] has found that the acoustic characteristics of ataxic dysarthric speech include slow speaking rate, mono-loudness and mono-pitch [14]. We have studied the acoustic characteristics of our collected data to see whether the characteristics match with the prior findings in the literatures.

5.1. Speaking Rate

We measure the speaking rate parameters from the passage reading (Task 3) and the counting task in Task 4 respectively. We calculate the average characters per second for each subject from Task 3 and the maximum characters per second from Task 4. The results are based on the second trial in both tasks. The results are shown in Figure 2. All dysarthric subjects (illustrated with white bubbles with horizontal lines) in Figure 2 have lower average and maximum syllables per second (x-axis) than control subjects (illustrated with gray bubbles). The maximum and average speeds are relatively consistent across different control subjects. In contrast, the maximum and average speaking rates vary substantially among the dysarthric subjects. The size of the bubbles represents the difference between maximum and average syllables per second. A large bubble size indicates the greater increase of the subject's speaking rate. The difference varies substantially across different dysarthric subjects. Many subjects cannot speak faster than normal conversational speed.

5.2. Loudness Control

We measure the intensity of speech produced by the dysarthric subjects and the controls from the Task 12, in which the subjects count from one to six in increasing loudness. The intensity is extracted using [17] by taking the maximum intensity (dB) from the vowel of each character in the second trial. Linear regression is applied to compute the rate of increase of the intensity of each word (as the slope of a straight line). We exclude the word "five" from the regression because in Cantonese, "five" ("五", /eng/) is a nasal syllable with significantly reduced intensity. Figure 3 shows the distributions of the voice intensity increase rate of the subjects and controls.

We observe that all control subjects can increase their loudness as instructed with positive slopes, but most dysarthric subjects have flat and even negative slopes. These subjects can only produce limited loudness differences. The results exhibit overall impairment in volume control.

5.3. Fundamental Frequency (F0) Control

Task 11 also provides us information about the ability of the subjects to control fundamental frequency (F0). In this task, the subjects sing the musical syllables "Do Re Mi Fa Sol La" in rising tones from the lowest pitch to the highest pitch. We measure the F0 (Hz) of each syllable and compute the difference (delta log F0) between the consecutive syllables (i.e. "Do Re", "Re Mi", etc.). The F0 is extracted using [17] taking the maximum pitch from the vowel of each syllable in the second trial. Figure 4 shows the distributions of the delta log F0 of the dysarthric subjects and the control speakers.

The control subjects can increase their F0 along the musical syllables as expected. Although most dysarthric subjects can also increase the F0 values as instructed, the degree of increase is less obvious than that of the control subjects. Sometimes the F0 may even drop as they advance up the scale. The results support the observation that the dysarthric subjects tend to have difficulty in pitch control.

6. Conclusions and Future Work

We have reported our current progress on developing a Cantonese dysarthric speech corpus. We have currently collected speech data from 11 Cantonese dysarthric subjects. The speech characteristics of the collected Cantonese dysarthric subjects match with the general descriptions of ataxic dysarthria. We will continue to collect more Cantonese dysarthric speech data to support our research. We will perform manual transcription, as well as the evaluation of speech intelligibility and severity of the dysarthric subjects. We believe that this corpus will be useful in supporting the study of Cantonese dysarthric speech and corresponding development of automation technologies.

7. Acknowledgement

The authors would like to thank Miss J Lee for her contribution in data collection process and transcription scheme. This project is partially sponsored by a grant from the Hong Kong SAR Government General Research Fund (reference no. GRF415513).

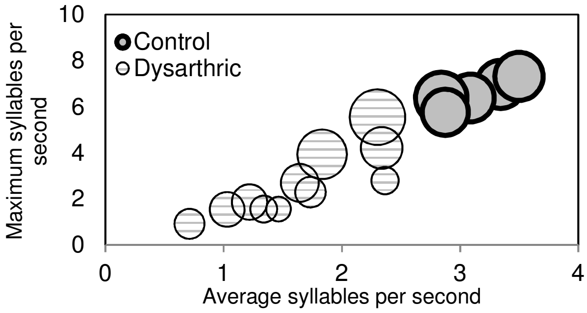

Figure 2: *The speaking rate of each subject is represented bv a bubble.*

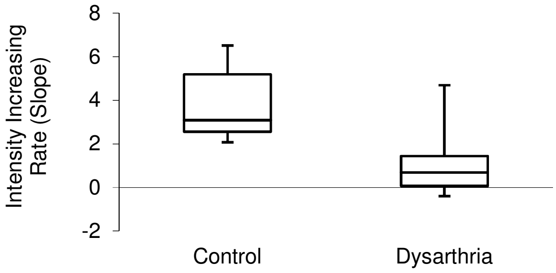

Figure 3: *The distribution of the rate of intensity increase of the subjects and controls. The rate indicates intensity (dB) change between consecutive numbers (one to six except five). The maximum and minium rate are shown as the top and bottom whiskers respectively.*

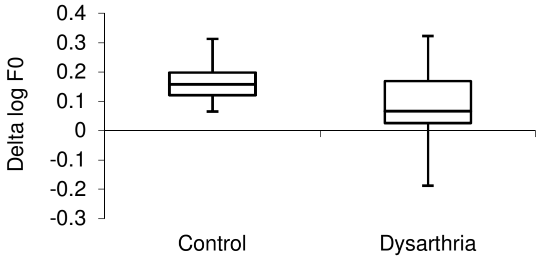

Figure 4: *The distribution of log F0 change.*

8. References

[1] W. K. L. Yam, H. S. S. Chan, K. W. Tsui, B. P. H. L. Yiu, S. S. L. Fong, C. Y. K. Cheng and C. W. Chan, "Prevalence Study of Cerebral Palsy in Hong Kong Children," *Hong Kong Medical Journal,* vol. 12(3), 2006.

[2] P. H. Chau, J. Woo, W. B. Goggins, Y. K. Tse, K. C. Chan, S. V. Lo and S. C. Ho, "Trends in Stroke Incidence in Hong Kong Differ by Stroke Subtype," *Cerebrovascular Disease,* vol. 31, pp. 138-46, 2011.

[3] S. Matthews and V. Yip, Cantonese: A Comprehensive Grammar (2nd edition), Routledge, 2011.

[4] X. Menéndex-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio and H. T. Bunnel, "The Nemours Database of Dysarthric Speech," in *International Conference on Spoken Language Processing*, 1996.

[5] H. J. 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.

[6] 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.

[7] P. M. Enderby, Frenchay Dysarthria Assessment, College Hill Press, 1983.

[8] C. S. Spillers, "The Rainbow Passage and The Grandfather Passage," [Online]. Available: http://www.d.umn.edu/~cspiller/readingpassages.html. [Accessed 21 January 2014].

[9] D. L. Choi, B. W. Kim, Y. J. Lee, Y. U. Lee, Y. N. Um and M. Chung, "Design and Creation of Dysarthric Speech Database for Development of QoLT Software Technology," in *O-COCOSDA*, 2011.

[10] 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.

[11] P. S. P. Cheung, A. Ng and C. K. S. To, "Hong Kong Cantonese Articulation Test," Language Information Sciences Research Centre, City University of Hong Kong, Hong Kong, 2006.

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

[13] P. Enderby, Frenchay Dysarthria Assessment-2, San Diego: College Hill Press, 2008.

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

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

[16] 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.

[17] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer [Computer Program]," 2014. [Online]. Available: http://www.praat.org. [Accessed 24 April 2014].