Development of a Cantonese Dysarthric Speech Corpus

Ka Ho Wong¹, Yu Ting Yeung², Edwin H. Y. Chan³, Patrick C. M. Wong⁴, Gina-Anne Levow⁵ and Helen Meng¹,²

¹ Human-Computer Communications Laboratory, ¹² Department of Systems Engineering and Engineering Management, ¹³ School of Life Sciences, ⁴CUHK-Utrecht University Centre for Language, Mind and Brain, ⁴Department of Linguistics and Modern Languages, ¹²,³,⁴ The Chinese University of Hong Kong, Hong Kong SAR, China ⁵Department of Linguistics, University of Washington, Seattle, WA USA

kh Wong@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 neurogenetic 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 re-
sults of dysarthric subjects are included. TORGO excludes 
non-words (e.g. “ah-p-e-ee”), short words and restricted sen-
tences.

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 dys-
arthric subjects and 20 male and 10 female non-dysarthric con-
trols. 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 
males and 10 females in Cantonese dysarthric subjects with cer-
bral palsy. The prompts include 100 monosyllable words such 
as /syu6/, “tree”, and /g4096/g3050/g18106 /gu1/ /ze1/ /hai2/ /bin1/ (“Aunt is where?”). Each initial, 
final, vowel and tone in Cantonese appeared a minimum of 
three times. However, the corpus is not publicly accessible. In 
additional to monosyllabic words, we also want to investi-
gate the speech production performance in different contexts.

3. Design of Cantonese Dysarthric Speech

3.1. Task Design

The stimuli we selected are structured to include a range of 
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 per-
formance characteristics such as the pitch control ability. The 
fifteen tasks performed by our subjects are as follows:
- 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 in-
itials and finals in the Jyutping system [12]. Task 1 con-
sists 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 sen-
tences are designed where the number of syllables of each 
sentence is similar to the sentence test in Frenchary Dys-
arthria Assessment (FDA-2) [13] and the sentences cover 
all Cantonese initials, finals and lexical tones. The aver-
age number of characters per sentence is 4.56 (standard 
development: 0.82).
- Task 3: Paragraph-level stimuli. The prompt of Task 3 is a 
phonetically rich passage widely used by Cantonese 
speech therapists1. 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 in-
cludes 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 character-
istics. For example, we follow the similar occurrence frequen-
cy of bilabial or labiodental phonemes as in the sentences of 
FDA-2. Table 2 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
synchronized between the notebook and the digital camera. Figure 1 illustrates the whole setup.

3.3. Data Collection

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, a reference video is 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.

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 shown in Figure 2. The results are shown in Figure 2. All dysarthric subjects (illustrated with grey bubbles). The maximum and average speeds are relatively consistent across different control subjects. In contrast, the maximum and aver-
age 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” (”/g7740”) 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).

8. References


