AUTOMATIC SPEECH RECOGNITION FOR ACOUSTICAL ANALYSIS AND ASSESSMENT OF CANTONESE PATHOLOGICAL VOICE AND SPEECH

Tan Lee[1,2]*, Yuanyuan Liu*[1,2]*, Pei-Wen Huang*[3]*, Jen-Tzung Chien*[3]*, Wang Kong Lam*[1]
Yu Ting Yeung[4]*, Thomas K.T. Law*[5]*, Kathy Y.S. Lee*[5]*, Anthony Pak-Hin Kong*[6]*, Sam-Po Law*[7]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong
[2]Language and Communication Disorder Laboratory, CUHK Shenzhen Research Institute
[3]Department of Electrical and Computer Engineering, National Chiao Tung University
[4]Stanley Ho Big Data Decision Analytics Research Centre, The Chinese University of Hong Kong
[5]Department of Otorhinolaryngology, Head and Neck Surgery, The Chinese University of Hong Kong
[6]Department of Communication Sciences and Disorders, University of Central Florida
[7]Division of Speech and Hearing Sciences, University of Hong Kong
tanlee@ee.cuhk.edu.hk

ABSTRACT

This paper describes the application of state-of-the-art automatic speech recognition (ASR) systems to objective assessment of voice and speech disorders. Acoustical analysis of speech has long been considered a promising approach to non-invasive and objective assessment of people. In the past the types and amount of speech materials used for acoustical assessment were very limited. With the ASR technology, we are able to perform acoustical and linguistic analyses with a large amount of natural speech from impaired speakers. The present study is focused on Cantonese, which is a major Chinese dialect. Two representative disorders of speech production are investigated: dysphonia and aphasia. ASR experiments are carried out with continuous and spontaneous speech utterances from Cantonese-speaking patients. The results confirm the feasibility and potential of using natural speech for acoustical assessment of voice and speech disorders, and reveal the challenging issues in acoustic modeling and language modeling of pathological speech.

Index Terms— Pathological speech, automatic speech recognition, acoustical analysis, objective assessment

1. INTRODUCTION

Speech is a preferred and natural modality of communication for human beings. Along the speech communication pathway, there are many imperfections that may obstruct the information flow. Among them, impairments on speech and language abilities are affecting the daily life of a large population worldwide. There is a strong demand for inter-disciplinary research and technology developments in the area of speech and language disorders and rehabilitation. The acoustic signal transmitted from the speaker to the listener provides an accessible and informative medium via which advanced signal processing and machine learning techniques can be applied to make contributions to improving speech communication.

The present study is focused on problems that are related to pathologies of voice and speech production. Two specific types of disorders are investigated: (1) **dysphonia**, and (2) post-stroke **aphasia**. Dysphonia, or voice disorder, is defined as "abnormality of pitch, volume, resonance and/or quality, and/or a voice that is inappropriate for the age, gender or culture of the speaker" [1]. Voice disorders are caused most commonly by the presence of irregular masses that affects normal vibration of the vocal folds or inefficient use of the laryngeal musculature. The irregular vibrations lead to abnormal perturbation to the glottal air flow and hence introduce atypical waveform changes in the acoustic signal. Aphasia refers to acquired language impairment resulting from a focal brain damage in the absence of sensory, motor, or cognitive impairments. Symptoms of aphasia can adversely affect one or more modalities of language. Language impairments associated with aphasia may be present across various linguistic levels. In terms of speech production, aphasia is often manifested by phonemic errors, articulation distortions, and speech disfluencies [2].

Assessment is a critical process in speech and language rehabilitation. It aims at determining the type or severity of impairment and identifying specific aspects of the disability. For dysphonia, perceptual evaluation of voice (PEV) is the standard procedure in clinical examination. It is carried out by trained speech pathologists, following a validated assessment protocol. The accuracy and reliability of PEV depend very much on the clinician's experience. For the assessment of aphasia, subjective evaluation of elicited oral discourses is a common practice. Comprehensive assessment of aphasia involves multi-level analysis of the oral discourses, which not only is time-consuming but also requires appropriate linguistic and cultural knowledge. There is clearly a strong desire to develop efficient and easy-to-use tools for objective assessment of speech and language disorders.

Acoustical analysis of speech has long been considered a promising approach to non-invasive objective assessment of voice and speech. In the past this required a lot of manual work on speech segmentation and feature extraction. Therefore the speech materials used for analysis were often limited to a small number of carefully elicited short utterances, e.g., sustained vowels, isolated words. With state-of-the-art spoken language technologies, it becomes possible to perform real-time analysis of natural speech and to automatically extract acoustic features from a large amount of heterogeneous data. This not only improves the efficiency and reliability of assessment but also leads to better user experience.

Speech and language pathology is an established profession in many Western countries. As the most populated country, China has been significantly left behind in this area and the demand for standardized assessment tools is highly pressing. However, linguistic differences do not allow straightforward adoption or migration of existing methods for Western languages, especially for language-related disorders like aphasia and dysarthria. This paper reports our effort toward acoustical analysis of natural Cantonese speech by dysphonia and aphasia patients. Cantonese is a major Chinese dialect spoken by tens of millions of people. Our research is based on two large-scale Cantonese speech databases, namely CanPEV and Cantonese AphasiaBank, which contain disordered voices and aphasia oral discourses respectively. We investigate the feasibility of applying the latest technology of automatic speech recognition (ASR) to free-content pathological speech, reveal the challenges in acoustic modeling and language modeling, and discuss the potential for clinical applications.

2. FUNDAMENTALS OF CANTONESE ASR

2.1. Characteristics of the Cantonese Dialect

Cantonese is spoken by tens of millions of people in the provinces of Guangdong and Guangxi, the neighboring regions of Hong Kong and Macau, and many overseas Chinese communities. In this paper, we focus on Cantonese as it is spoken in Hong Kong. Like in Putonghua or Mandarin, Cantonese is a monosyllabic and tone language. Each Chinese character is pronounced as a monosyllable carrying a specific lexical tone. It is the smallest meaningful unit (morpheme) of the language. A character may have several different pronunciations. A tonal syllable in Cantonese may also correspond to multiple characters. A spoken word is composed of one or more syllables and a spoken sentence is a sequence of continuously uttered syllables.

Each Cantonese syllable is composed of two parts: the *Initial* (onset), and the *Final* (rime). The *Initial* is typically a consonant, while the *Final* contains a vowel nucleus followed by a consonant coda. The initial consonant and the coda are optional. There are 20 *Initials* and 53 *Finals* in Cantonese, which lead to over 600 legitimate *base syllables*. Each *base syllable* can be associated with different tones. If the tone is changed, the syllable generally refers to another character that has a different meaning. Traditionally, Cantonese is said to have nine tones. Three of them are known as the entering tones and the others are non-entering tones. Entering tones are contrastively shorter than non-entering tones. In terms of pitch level, each entering tone coincides with one of the non-entering tones. Therefore it is common to define six distinctive tones of Cantonese [3].

2.2. Cantonese ASR System

The key components of a large vocabulary continuous speech recognition (LVCSR) system are acoustic models, pronunciation lexicon, and language models. The acoustic models are typically a set of hidden Markov models (HMMs) that characterize the statistical variation of input speech features. For Cantonese speech recognition, *Initials* and *Finals* are commonly used as the basic units for acoustic modeling. Context-dependent modeling is applied to form more specific units based on the phonetic context. Each of the context-dependent *Initials* and *Finals* is represented by a dedicated HMM with 3 to 6 states [3]. For the training of Cantonese acoustic models, a large-scale speech database named CUSENT was developed at the Chinese University of Hong Kong (CUHK)[4].

The pronunciation lexicon and the language models are application and domain dependent. The pronunciation lexicon lists all syllables and/or words being used in the target application and specifies the constituting *Initials* and *Finals*. The language models give the probabilities of the occurrences and co-occurrences of the lexical entries. A search algorithm is used to determine the most likely syllable/word sequence, given an input sequence of acoustic feature vectors. The search space is constructed by connecting the HMM states in the *Initial* and *Final* models according to the pronunciation lexicon and the language models.

In this study, a baseline ASR system for Cantonese is developed using the Kaldi speech recognition toolkit [5]. Two different methods of acoustic modeling are investigated:

GMM-HMM: The acoustic feature vector is computed with a context window of 7 frames, each being represented by 13 MFCC features. Linear discriminant analysis (LDA) is applied to project the contextual feature vector into 40 dimensions, followed by the maximum likelihood linear transform (MLLT). Speaker adaptive training (SAT) is performed on both training and test utterances by using the feature-space maximum likelihood linear regression (fMLLR) transform. HMMs are trained to model the 20 *Initials* and 53 *Finals*. Each HMM has 3 emission states. The training data consist of $20,378$ utterances of read-style continuous speech from CUSENT. The technique of decision-tree state tying is used for context-dependent modeling. As a result, $2,288$ context-dependent HMM states (senones) are obtained and they are represented by $24,023$ Gaussians.

DNN-HMM: The acoustic feature vector is composed of 40-dimensional fMLLR features from a context window of 11 frames. The same HMM topology as in the GMM-HMM system is adopted, except that a deep neural network (DNN) is used to generate the state-level posterior probabilities. The DNN contains 6 hidden layers and each layer has $1,024$ neurons. The number of output neurons is $2,288$, i.e., equal to the number of HMM states. The restricted Boltzmann machine (RBM) is used to initialize the neural network parameters and subsequent training is done by the back-propagation algorithm via stochastic gradient descent.

The baseline performance of the GMM-HMM and DNN-HMM acoustic models is evaluated with $1,198$ test utterances from CUSENT. The pronunication lexicon contains 640 Cantonese base syllables. The language model is a uniform syllable uni-gram, i.e., all syllables are assumed to be equally probable. The syllable error rates produced by the GMM-HMM and DNN-HMM systems are 15.6% and 10.1%, respectively. The superior performance of DNN-HMM with respect to GMM-HMM is in agreement with experimental results on other languages.

3. PATHOLOGICAL SPEECH DATABASES

3.1. CanPEV

The MEEI database is by far the most commonly used database of disordered voice [6, 7]. It contains $1,400$ voice samples, including sustained vowels and short passages in American English, from speakers with a wide variety of voice disorders. Voice databases of similar scale are rarely seen for other languages. CanPEV is a Cantonese voice database developed by the Division of Speech Therapy of the Chinese University of Hong Kong (CUHK). It was intended initially for perceptual training of speech therapists on voice disorder. The entire database contains speech recordings from 232 subjects with normal or pathological voices. All subjects are native speakers of Cantonese. The speech data from each subject are divided into the following three parts.

Sustained vowels: Three repetitions for each of the vowels /a/, /i/, /u/. Each vowel is about 3-5 seconds long;

Passage reading: Reciting of a given passage that describes Hong Kong. The passage contains 146 Chinese characters. The duration is about 40 second each passage;

Conversational speech: Spontaneous responses to the questions "*What have you been doing today?*" and "*Can you comment on your own voice?*"

All recordings were made in quiet environment with sampling frequency of $44,100$ Hz with 16-bit linear quantization. Perceptual assessment on the voice data in CanPEV was performed by a group of experienced speech therapists. The speech therapists were asked to listen to all speech materials from each subject and to rate the voice on *overall severity* and a number of specific vocal parameters, e.g., *roughness*, *breathiness*, *strain*. The ratings were given on a 10-point scale. The overall rating was obtained by averaging the ratings from the participating listeners.

3.2. Cantonese AphasiaBank

There have been very few studies on acoustical analysis of aphasia speech for languages other than English. The absence of properly collected and annotated speech corpora for the target language is a major challenge. Since 2009, a large-scale project on multi-modal and multi-level examination of Chinese aphasic discourse has been carried out jointly at the University of Central Florida and the University of Hong Kong. As the core part of this project, the **Cantonese AphasiaBank** was established to support both fundamental and clinical research on Cantonese-speaking aphasia population [8].

Cantonese AphasiaBank contains audio recordings of spontaneous oral narratives from 149 unimpaired native Cantonese speakers and 104 individuals with post-stroke aphasia. The speech materials were elicited using the AphasiaBank protocol [9], with adaptation to local Chinese culture. There were 8 elicitation tasks for each speaker, including picture descriptions, procedure description, story telling and monologue. Audio signals were recorded using a head-worn condenser microphone and a digital recorder with $44,100$ Hz sampling. All recordings were manually transcribed using the Child Language ANalyses computer program [10]. The orthographic transcription of each recording is in the form of a sequence of Chinese characters.

The subjects with aphasia went through a standard assessment process using the Cantonese Aphasia Battery [11]. Their severity of aphasia was reflected by the Aphasia Quotient (AQ), which is a composite value based on their performance in multiple language tasks. The maximum value of AQ is 100 and a lower value indicates a higher severity level of aphasia.

4. EXPERIMENTS AND DISCUSSION

4.1. ASR on Disordered Voice

4.1.1. ASR Results

Based on the perceptual ratings on overall severity, the subjects in CanPEV were divided into 4 categories: *normal*, *mild*, *moderate* and *severe*. ASR experiments are carried out with the "passage reading" speech of 10 subjects from each of the *mild*, *moderate* and *severe* categories. The acoustic models, the pronunciation lexicon and the language model are the same as in the baseline Cantonese ASR system described in Section 2.2. Table 1 shows the speech recognition performance in terms of syllable error rates for the three categories of voice data.

4.1.2. Discussion

The speaking style of "passage reading" utterances in CanPEV is similar to the test data of CUSENT. With the uniform syllable uni-gram, the results in Table 1 mainly reflect the degree of acoustic model mismatch that is caused by the change of voice quality. The speech recognition accuracy shows a clear trend of declining from the *mild* category to *moderate* and *severe*. This seems to suggest that the recognition accuracy itself could be a good assessment metric. The superiority of DNN-HMM over GMM-HMM is noticeable for CanPEV data. For the *mild* category, the syllable error rate is close to that on CUSENT (14.6% vs. 10.1%). The performance advantage of DNN-HMM becomes less significant with severity level increasing. More effective training strategies are needed to model disordered voice.

Table 1. Syllable error rates on "passage reading" speech of CanPEV

	Mild	Moderate	Severe
GMM-HMM	24.1%	36.9%	56.2%
DNN-HMM	14.6%	28.1%	48.8%

For acoustical voice assessment, the goal of research has been to identify useful acoustic parameters that can characterize waveform and spectral irregularities [12, 13]. Previous studies were mostly limited to sustained phonation data. Although the phonetic homogeneity allows the investigation to be more focused, many voice problems might not be revealable in isolated vowel sounds. It was shown that segmental and suprasegmental linguistic factors of connected speech, especially at consonant-vowel transitions, had strong influence on voice quality [14]. It was also found that perceptual assessment using connected and conversational speech was more reliable than sustained vowels [15]. For practical applications, the use of free-content natural speech is preferable, as it represents the real speech communication for humans.

The results of ASR experiments on CanPEV data show the potential of using natural continuous speech for acoustical voice assessment. Despite the relatively low recognition accuracy for the *severe* category, the acoustic models are at least sufficient for automatically segmenting and identifying target phonemes in broad classes, e.g., vowels. A detailed analysis of the substitution errors shows that most of the confusions are between syllables sharing similar vowel nuclei.

4.2. ASR on Aphasia Speech

4.2.1. ASR Results

There are 8 elicitation tasks for each subject in the Cantonese AphasiaBank. Speech recognition experiments are carried out with the recordings from one of the tasks, which is a sequential picture description about a boy breaking a window by accident. A total of 17 aphasia speakers and 17 age- and gender-matched unimpaired speakers are selected for the experiment. The 17 aphasia speakers include 14 Anomic, 2 Transcortical Sensory and 1 Wernickes aphasia. The AQ of these aphasia patients range from 73.2 to 99.0 (full score: 100).

The original orthographic transcriptions of the selected recordings were manually checked and converted into Cantonese syllable pronunciations. The speech data from aphasia and normal speakers contain $1,194$ and $1,699$ syllables, respectively. The acoustic models are the same as in the baseline Cantonese ASR system as described in Section 2.2 . The pronunciation lexicon contains about 640 base syllables. The language model is a task-specific syllable uni-gram model trained with the transcriptions of the recordings of the same elicitation task from all unimpaired speakers in the database. Table 2 shows the speech recognition performance for the 17 aphasia and 17 unimpaired speakers.

Table 2. Syllable error rates on spontaneous speech from Cantonese AphasiaBank

	Aphasia	Unimpaired
GMM-HMM	58.2%	43.9%
DNN-HMM	57.8%	42.7%

4.2.2. Discussion

In our previous work [16], acoustical analysis of Cantonese aphasia speech was carried out based on HMM forced alignment results. It was shown that aphasic speech exhibits distinctive characteristics in supra-segmental duration, which demonstrate good potential for assessment purpose. The present study is taking one step further toward fully automated analysis of natural speech. Due to the spontaneous nature of speech, high error rates are observed for the Cantonese AphasiaBank. Even for unimpaired speakers, the syllable error rate is over 40% with GMM-HMM. The advantage of DNN-HMM is also not as significant as that on CanPEV. This may imply that acoustic models are not the most critical factor causing the low accuracy. Speaking style mismatching and the lack of good language models are the challenges. Indeed, the natural discourses of picture description contain frequency occurrences of non-speech sounds and filler words, in the cases of both aphasia and normal speech. These events are not properly modeled in the current version of our ASR system. On the other hand, building domain-specific language models would be helpful to improve the ASR performance. This is practically feasible and appropriate since aphasia assessment is done with a few standardized narrative tasks.

ACKNOWLEDGEMENT

This research is partially supported by a GRF project grant (Ref: 14204014) from Hong Kong Research Grants Council and by the Shenzhen Municipal Engineering Laboratory of Speech Rehabilitation Technology. The CanPEV database was developed with the support of GRF project (Ref: 468708). The Cantonese AphasiaBank was supported by a grant funded by the National Institutes of Health (NIH-R01-DC010398).

5. REFERENCES

[1] Sylvia Taylor-Goh, "RCSLT Clinical Guidelines," 2005.

[2] Julie L Wambaugh, Patrick J Doyle, Michelene M Kalinyak, and Joan E West, "A critical review of acoustic analyses of aphasic and/or apraxic speech," *Clinical Aphasiology*, vol. 24, pp. 35–64, 1996.

[3] P.C. Ching, Tan Lee, W. K. Lo, and Helen Meng, "Cantonese speech recognition and synthesis," in *Advances in Chinese Spoken Language Processing*, C.-H. Lee et al., Ed., pp. 365–386. World Scientific Publishing, Singapore, 2006.

[4] Tan Lee, W. K. Lo, P. C. Ching, and Helen. Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, vol. 36, pp. 327–342, 2002.

[5] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," 2011.

[6] "Elemetrics disordered voice database (version 1.03)," 1994, Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory, Boston, MA.

[7] "Disordered voice database and program, model 4337," `http://www.kayelemetrics.com/`.

[8] Anthony Pak-Hin Kong, Sam-Po Law, and Alice Lee, "The construction of a corpus of cantonese-aphasic-discourse: a preliminary report," in *American Speech-Language-Hearing-Association Convention*, 2009.

[9] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[10] Brian MacWhinney, "The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database," *Computational Linguistics*, vol. 26, no. 4, pp. 657–657, 2000.

[11] Edwin ML Yiu, "Linguistic assessment of chinese-speaking aphasics: Development of a cantonese aphasia battery," *Journal of Neurolinguistics*, vol. 7, no. 4, pp. 379–424, 1992.

[12] Juan Ignacio Godino-Llorente, Pedro Gómez-Vilda, and Tan Lee, "Analysis and signal processing of oesophageal and pathological voices," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 12, 2009.

[13] Maria Markaki and Yannis Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.

[14] Anders Löfqvist and R.S. McGowan, "Voice source variations in running speech," in *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, J. Gauffin and B. Hammarberg, Eds., pp. 113–120. San Diego, CA, 1991.

[15] Thomas Law, Jean H Kim, Kathy Y Lee, Eric C Tang, Joffee H Lam, Andrew C van Hasselt, and Michael C Tong, "Comparison of raters reliability on perceptual evaluation of different types of voice sample," *Journal of Voice*, vol. 26, no. 5, 2012.

[16] Tan Lee, A Kong, V Chan, and Haipeng Wang, "Analysis of auto-aligned and auto-segmented oral discourse by speakers with aphasia: A preliminary study on the acoustic parameter of duration," *Procedia-Social and Behavioral Sciences*, vol. 94, pp. 71–72, 2013.