INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Acoustic Assessment of Disordered Voice with Continuous Speech Based on Utterance-level ASR Posterior Features

Yuanyuan Liu[1,2], *Tan Lee*[1,2], *P.C. Ching*[1], *Thomas K.T. Law*[3], *Kathy Y.S. Lee*[2,3]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK)
[2]Language and Communication Disorder Laboratory, CUHK Shenzhen Research Institute
[3]Department of Otorhinolaryngology, Head and Neck Surgery, CUHK
yuanliu@ee.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

Most previous studies on acoustic assessment of disordered voice were focused on extracting perturbation features from isolated vowels produced with steady-state phonation. Natural speech, however, is considered to be more preferable in the aspects of flexibility, effectiveness and reliability for clinical practice. This paper presents an investigation on applying automatic speech recognition (ASR) technology to disordered voice assessment of Cantonese speakers. A DNN-based ASR system is trained using phonetically-rich continuous utterances from normal speakers. It was found that frame-level phone posteriors obtained from the ASR system are strongly correlated with the severity level of voice disorder. Phone posteriors in utterances with severe disorder exhibit significantly larger variation than those with mild disorder. A set of utterance-level posterior features are computed to quantify such variation for pattern recognition purpose. An SVM based classifier is used to classify an input utterance into the categories of mild, moderate and severe disorder. The two-class classification accuracy for mild and severe disorders is 90.3%, and significant confusion between mild and moderate disorders is observed. For some of the subjects with severe voice disorder, the classification results are highly inconsistent among individual utterances. Furthermore, short utterances tend to have more classification errors.

Index Terms: disordered voice, continuous speech, speech recognition, acoustic posteriors

1. Introduction

Speech disorder is also known as speech impairment. People with speech disorders are not able to pronounce sounds correctly, speak fluently and naturally, or sometimes have problems with their voices. These problems may have significant negative impact on the daily life of impaired individuals and create low self-esteem. As a major type of speech disorder, voice disorder refers to a variety of abnormality in vocal quality, pitch, loudness, resonance, and/or duration, which generally do not depend on the spoken content [1]. Organic voice disorders result from alterations in respiratory, laryngeal, or vocal tract mechanisms. Functional voice disorders are caused by improper or inefficient use of the vocal mechanism, which may be related to occupation or personal habits.

Comprehensive assessment of voice disorder is carried out with a combination of methods, including laryngeal imaging, perceptual evaluation and acoustic analysis. Among them, non-invasive perceptual evaluation is most commonly used in everyday clinical practice. It is done by asking the patient to produce a set of designated speech sounds, which are rated by a trained speech therapist based on auditory perception. This approach is known to be subjective. The accuracy and reliability of assessment depend greatly on the type of speech being rated and the therapist's experience. Acoustic analysis of speech signal has long been considered a promising approach to objective assessment of disordered voice. It is done by observing and quantifying voice-related signal properties, with reference to normal voices.

Previous studies on acoustic assessment of disordered voice were focused mainly on extracting perturbation features from isolated vowels produced with steady-state phonation. Natural continuous speech, however, is more preferable for various reasons. First, the ecological validity can be ensured since continuous speech is used in real-world situations [2][3]. The research findings are expected to be more applicable to practical work. Second, voice quality is highly influenced by linguistic variation. There are voice problems that cannot be elicited with sustained vowels [4]. In [5], it was found that auditory-perceptual evaluation with natural speech had higher level of inter-rater and intra-rater reliability than sustained vowels.

Fully automatic assessment of voice with natural speech meets many technical challenges [6]. With highly varying linguistic content, the commonly used long-term features, e.g., F0, intensity, jitter, shimmer, would lose their effectiveness. It is believed that certain phonemes are particularly useful for detecting voice problems. Identifying and locating these target phonemes in natural speech are not trivial tasks. The short phoneme duration also makes it difficult to extract effective voice features. The present study aims to develop an ASR based approach to automatic assessment of disordered voice with natural speech utterances.

In [2], an automatic system was developed to assess the speech from patients suffering neck and head tumors. The system extracted sentence-level acoustic features from manually labeled continuous speech. All vowel segments in an utterance were concatenated to facilitate computation of long-term features like jitter, shimmer and harmonic-to-noise ratio. The same problem of speech intelligibility assessment was tackled in [7]. It first computed low level features at frame-level, e.g. delta MFCCs and formant frequencies, so a set of feature vectors were obtained for each utterance. Then three different types of speech representations (i.e., linear subspace, covariance matrix and Gaussian distribution) were exploited on the feature vectors of each utterance. To measure the similarity and difference among samples, various kernels were applied to these representations. In [8], diverse types of speech were used for determining the severity of Parkinson's disease. Apart from computing conventional perturbation and spectral features at frame-level, the entropy of phone posteriors distribution was measured for each time frame, where a triphone acoustic model in ASR system was used to generate phone posteriors for each frame in an utterance. To transfer frame-level features to sentence-level

http://dx.doi.org/10.21437/Interspeech.2017-280

features, i-vector and statistical functions were utilized.

In our recent work [9], an initial attempt was made toward acoustical assessment of voice with continuous speech from Cantonese-speaking patients. Disordered speech utterances were processed by a large-vocabulary ASR system trained with normal speech. Not surprisingly the speech recognition accuracy declined substantially as disorder severity increased. It was revealed that phone posterior probabilities retrieved from the DNN-HMM based ASR system could potentially be used as a discriminative indicator of the severity level. A set of most discriminative phones (mostly voiced sounds) were identified. These findings motivated the present investigation on automatic voice assessment. As shown in Figure 1, a pattern classifier is used to classify the input utterances into three different categories, namely **mild**, **moderate** and **severe**. The input features for classifier are directly obtained from or derived based on the ASR system output in response to the input utterance. We aim at utterance-level assessment and each utterance contains a continuously spoken sentence. If there are multiple utterances available from a subject, the subject-level assessment can be done by combining the utterances-level results.

In the next section, the Cantonese ASR system is described and the ASR posterior features obtained from speech utterances of different severity levels are analyzed and compared. The design of utterance-level posterior features and the experimental results of SVM based classification of voice disorder severity are presented in Section 3. Discussion and conclusion will be given in Section 4 and 5 respectively.

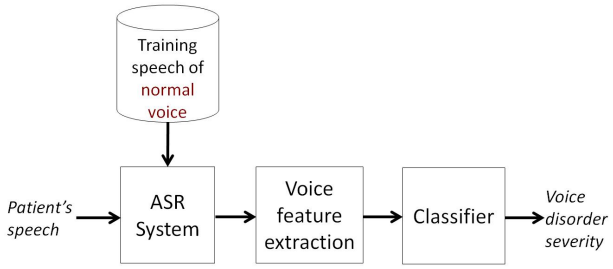

Figure 1: *Automatic classification of disordered voice.*

2. Analysis of acoustic posteriors

2.1. Cantonese speech database

Cantonese is a major Chinese dialect used in Southern China. Like other Chinese dialects, Cantonese is a monosyllabic and tonal language. Each Chinese character is pronounced as a monosyllable sound with a specific tone. Cantonese syllables are often described by the Initial-Final structure [10]. The Initial is a consonant, which could be voiced or unvoiced. The Final consists of a vowel nucleus followed by an optional coda. There are a total of 13 vowels and 19 consonants in Cantonese, from which over 600 legitimate base syllables are formed. Each base syllable can be associated with a number of different tones [10].

CUSENT is a large-scale continuous speech database of Cantonese developed at the Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK) [11]. CUSENT contains about $21,200$ utterances from 80 native speakers with normal voices. The spoken content covers over $5,000$ newspaper sentences.

CanPEV is a voice database developed at the Division of Speech Therapy of CUHK [12]. The database contains different types of speech from 232 subjects with normal or pathological voices. All subjects are native speakers of Cantonese. The speech data from each subject consist of three parts: sustained vowels, passage reading and spontaneous speech. A group of trained speech therapists performed perceptual rating of voice on each subject. Numerical ratings in 10-point scale were given on the overall severity of voice disorder and specific voice problems (e.g., roughness, breathiness, strain). Based on the numerical scores, the subjects are divided into 4 categories, namely **normal**, **mild**, **moderate** and **severe**. The present investigation is focused on the passage-reading part of CanPEV. The content of speech is the same for all subjects, i.e., a short passage of describing Hong Kong. The duration of reading speech is around 40 seconds for each subject.

2.2. Cantonese ASR system

The Kaldi Speech Recognition Toolkit [13] was used to develop a DNN-HMM based ASR system for Cantonese. For acoustic model training, about $20,000$ utterances from CUSENT and 429 utterances from normal speakers in CanPEV were used. The input feature vector contains 440-dimension fMLLR features, covering 11 contextual frames. The DNN has 6 hidden layers with the size of 1024. The output layer has $2,437$ neurons, each outputting a posterior probability of a specific distinct triphone state. The DNN was initialized with restricted Boltzmann machine (RBM) and trained using the back-propagation algorithm via stochastic gradient descent.

The pronunciation lexicon is composed of 630 base syllables of Cantonese. Given an input utterance, the ASR system generates a sequence of base syllables using uniform syllable uni-grams, i.e., all syllables are assumed equally probable. In this way, the ASR performance reflects mainly the acoustic mismatch caused by voice disorders. The passage-reading utterances from 30 speakers of CanPEV are used as the ASR test data. There are 10 speakers in each of **mild**, **moderate** and **severe** overall severity categories. Each speaker has 13 test utterances. The syllable error rates (SER) on the three categories are 7.76%, 17.74% and 43.08%, respectively.

2.3. Posterior probabilities in utterances

Frame-level phone posterior probabilities can be derived from the DNN softmax layer. At each time frame, the softmax layer outputs a $2,437$-dimension vector of posterior probabilities. Each element of the vector corresponds to a distinct triphone state. For each of the 32 phones (13 vowels and 19 consonants), the frame-level posterior is computed by summing up the state posteriors that are associated with this phone.

In our previous study [9], a strong correlation between phone posteriors and perceptual ratings of voice disorder was revealed. The posterior probabilities of vowels and voiced consonants were found to be more discriminative than those of unvoiced consonants. By analyzing the phone posteriors over different severity categories, we identify 5 discriminative vowels that frequently occur in the passage-reading part of CanPEV. They are: /a/, /i:/, /I/, /O:/, /u:/ (/ʌ/, /i:/, /ɪ/, /ɔ:/, /u:/ in IPA) [1]. In the following let us examine the frame-level phone posteriors of these vowels in a pair of example utterances. One of the utterances is from the **mild** category and the other from **severe**.

[1] In this paper, the SPPAS-dict symbols are used to represent Cantonese phonemes [14]. The mapping between SPPAS-dict and IPA can be found at http://dsp.ee.cuhk.edu.hk/research-tools/Cantonese-phones.html

Transcription

Chinese character	才能争取到今天的成就。
Syllables	coi nang zang ceoi dou gam tin dik sing zau.
Phones	ts_h O: i: n a N ts a N ts_h eo i: t o u: k a m t_h i: n t l k s l N ts a u:

ASR output

mild	coi nang zang ceoi dou gam tin dik sing zau
severe	coi leon zang ceoi dou gam tin dik caa zang

Figure 2: *Transcriptions and ASR outputs for an example utterance.*

They have the same spoken content as given in Figure 2. The ASR outputs for the two utterances are also shown in the figure [2]. As expected, the ASR output for the **severe** utterance contains more errors (underlined phones) than the **mild** one.

Figure 3 shows the evolution of frame-level phone posteriors over the **mild** utterance. There are five plots, each corresponding to one of the five vowels. The horizontal axis indexes time in frames and the vertical axis represents the posterior probability. The ASR output is also shown in a time-aligned manner with the posterior curves. At the time interval where a particular vowel is present (and recognized), the respective posterior probability shows a large value, i.e., close to 1. Otherwise the posterior is close to 0. Typically only one phone with dominantly high posterior can be identified at any specific time. For example, /O:/ and /i:/ in the syllable *coi* are present with high posterior probabilities at frames $75-85$ and $85-90$, respectively.

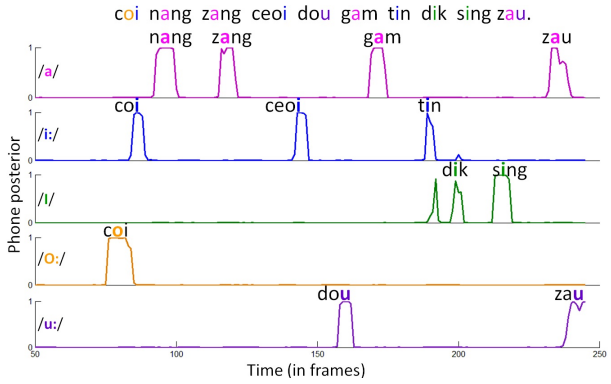

Figure 3: *Phone posteriors in the mild utterance.*

Similar posterior plots for the **severe** utterance are given in Figure 4. Compared to Figure 3, the phone posteriors exhibit significantly greater variability in the **severe** utterance. It may not be straightforward to identify a clearly "dominating" phone that dominates in posterior probability. For the correctly recognized phones, the posterior values could be much lower than 1. Amplified plot of posteriors of /a/ in the **severe** utterance is shown in Figure 5, where the variability is displayed clearly.

The above observations suggest that the variability of phone posteriors in continuous speech is a good indicator of voice disorder severity. In the next section, a novel approach to utterance-level voice assessment will be developed by exploiting the variability of phone posteriors.

[2]Cantonese syllable transcriptions are romanised using the JyutPing system [10] . The mapping between JyutPing and IPA can be found at http://dsp.ee.cuhk.edu.hk/research-tools/Cantonese-phones.html

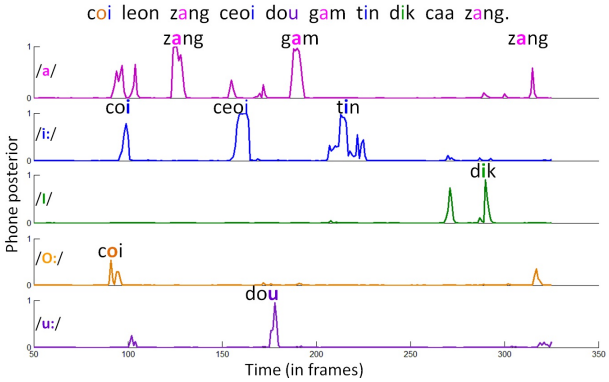

Figure 4: *Phone posteriors in the severe utterance.*

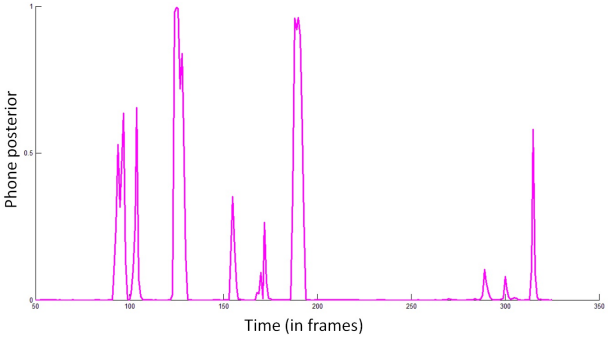

Figure 5: *Posteriors of /a/ in the severe utterance.*

3. Classification of disorder severity

As shown in Figure 1, voice assessment is formulated as a pattern classification problem. For each input utterance, which typically contains a continuously spoken sentence, a set of posterior-based features are derived from the ASR output. Based on the features, the utterance is classified into three classes of disorder severity: **mild**, **moderate** and **severe**.

3.1. Utterance-level posterior features

As seen in Section 2.2, the accuracy of ASR varies substantially as the severity of disorder changes. The recognized syllable sequence (or phoneme sequence) contains an unknown number of errors and thus should not be used directly for feature extraction purpose. We need a robust method of extracting utterance-level features that does not rely on the ASR performance. The examples in Section 2.3 show that frame-level phone posteriors in **mild** utterances tend to be concentrated around the extremity values, whereas those in **severe** utterances may fluctuate in a wide range. Specifically, a recognized phone in a **mild** utterance has a clearly dominating posterior value, i.e., very close to 1. For a **severe** utterance, the "dominating" phone may not have a very high posterior because of the acoustic mismatch.

For each of the five selected vowels, we try to locate all time frames where the vowel is likely to be recognized and compute the variation of posteriors in these frames. An utterance-specific threshold is determined as the 97% percentile of the frame-level posteriors of this vowel in the utterance. The time frames with posterior values above this threshold are identified as the "recognized" frames and the standard deviation of these frame-level posteriors is computed as an utterance-level feature for classification. As a result, each input utterance is represented by a

5-dimension feature vector, which is composed of the posterior variation features of the five vowels.

The use of a percentile threshold is based on the assumption that the target vowel occupies a certain percentage of time in the utterance. The exact percentage can never be determined accurately. In the case that the target vowel is not present in an utterance, the frame-level posteriors would all be close to 0 and the computed feature of posterior variation does not depend much on the threshold. On the other hand, if the target vowel occupies a long duration or occurs repetitively in an utterance, only the "best-matched" frames would be retained with the use of percentile threshold. In the present study, the threshold of 97% is determined empirically.

3.2. Two-class classification

A support vector machine (SVM) aims to separate two classes of patterns with an optimal hyperplane. It is used to perform classification between each pair of severity categories, namely **mild-severe**, **mild-moderate** and **moderate-severe**. In each case, there are about 260 utterances, i.e., 130 from one of the two categories. Classification experiments are carried out with the arrangement of 5-fold cross validation. The "quadratic" SVM kernel is used. The input feature vector consists of the 5 posterior deviation features as described in Section 3.1. The classifier output is the recognized category of disorder severity.

The error rate for **mild-severe** classification is 9.7% (or accuracy of 90.3%). There are 8 **mild** utterances being mis-classified as **severe** and 17 **severe** utterances mis-classified as **mild**. For **moderate-severe** classification, the error rate is 10.9% and the majority of classification errors are from the **severe** class. The most confusing pair of categories is **mild-moderate**, with an error rate of 42.7% and the errors are about evenly distributed in both classes.

3.3. Three-class classification

A three-class model is implemented by training 3 binary classifiers and applying the one-versus-one strategy. All of the binary classifiers have the same attribute setting. The input feature vector is assigned to the class with the least binary loss. Table 1 shows the confusion matrix of classification. The recalls for **mild**, **moderate** and **severe** classes are 60.0% (78/130), 53.1% (69/130) and 82.7% (105/127), respectively.

Table 1: *Confusion matrix for the mild, moderate and severe*

perceptual label	classification result		
	mild	moderate	severe
mild	78	42	10
moderate	56	69	5
severe	9	13	105

4. Discussion

The experimental results show that the proposed method is very effective in discriminating **severe** disorder from **mild** and **moderate** disorders. The classification accuracies of the **mild-severe** pair and the **moderate-severe** pair are both around 90%. Table 1 shows that, among the 260 **mild** and **moderate** utterances, only 15 are mis-classified as **severe** disorder, while 22 of the 127 **severe** utterances are mis-classified as **mild** or **moderate**. However, it seems to be very difficult to distinguish between **mild** and **moderate** disorders.

There are more cases of **severe** utterances being mis-classified as **mild** or **moderate** than the opposite direction. It should be noted that the perceptual evaluation for CanPEV data was done on per subject basis. In other words, all utterances from the same subject have the same ground-truth class label even if the voice characteristics may vary significantly across individual utterances. The experimental results suggest that the disorder characteristics of the same **severe** subject may be localized in some of his/her utterances and not appear in the others. The existence of these **mild** (or even **normal**) utterances could not be reflected in the subject-level rating. On the contrary, for a subject to be rated as **mild**, it is required that most, if not all, of his/her utterances do not show any disorder symptom.

It is noted that two of the **severe** subjects have half of their utterances classified as **mild**. For these two speakers, the ASR performances in terms of phone accuracy are 92.7% and 97.3% respectively, meaning that there is no significant mismatch between the MFCCs extracted from these speakers' voices and those from normal voices. Interestingly both speakers were rated to be **severe** in terms of roughness of voice. While roughness is a distinctive perceptual feature of disordered voice, it is not effectively represented in MFCCs. Further investigation is needed to search for alternative or supplementary acoustic features.

Among the 13 utterances from each speaker, the one with the shortest sentence length has the most classification errors. As a matter of fact, there are only two occurrences of the target vowels in this short utterance, which may not be adequate for measuring the disorder severity. On the other hand, in the present study, the training data and test data from all speakers have the same linguistic content. However, since the posterior features are extracted from a set of commonly used vowels, the proposed method is expected to perform reasonably well when the spoken content is changed, as long as the target vowels are present in the input utterances.

5. Conclusions

A preliminary framework of automatic voice assessment using natural speech has been described. It is based on a general-purpose ASR system trained with normal voice. The ASR system has two major functions in the design. First, it performs phoneme recognition on input speech and hence helps locating specific phonemes for voice feature extraction. Second, the ASR posterior probabilities are effective features for voice assessment. The proposed method is able to differentiate **severe** disordered voices from **moderate** and **mild** ones, while the confusion between **mild** and **moderate** remains significant. The current system design is basic and flexible. It can be improved in several aspects, e.g. introducing articulatory features, exploiting interpretable acoustic features of voiced segments based on ASR output, applying semi-supervised learning with more training data and advancing to subject-level voice assessment by fusing utterance-level predictions. The system design is also applicable to other languages and generalizable to spontaneous speech.

6. Acknowledgements

This research is partially supported by a GRF project grant (Ref: 14204014) from Hong Kong Research Grants Council, and the Major Program of National Social Science Fund of China (Ref: 13&ZD189), and by the Shenzhen Municipal Engineering Laboratory of Speech Rehabilitation Technology.

7. References

[1] S. L. H. Association, "Definitions of communication disorders and variations," 1993.

[2] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," vol. 29, no. 1, 2015, pp. 132 – 144.

[3] M. B. Brewer, H. Reis, and C. Judd, "Research design and issues of validity." New York, NY, US: Cambridge University Press, xii, 558 pp., 2000, pp. 3–16.

[4] A. Löfqvist and R. 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., San Diego, CA, 1991, pp. 113–120.

[5] T. Law, J. H. Kim, K. Y. Lee, E. C. Tang, J. H. Lam, A. C. van Hasselt, and M. C. Tong, "Comparison of rater's reliability on perceptual evaluation of different types of voice sample," in *Journal of Voice*, vol. 26, no. 5, 2012, pp. 666.e13 – 666.e21.

[6] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions," in *ICASSP*, 2016, pp. 6470–6474.

[7] D. Y. Huang, M. Dong, and H. Li, "Combining multiple kernel models for automatic intelligibility detection of pathological speech," in *ICASSP*, 2016, pp. 6485–6489.

[8] J. Kim, M. Nasir, R. Gupta, M. V. Segbroeck, D. Bone, M. Black, Z. I. Skordilis, Z. Yang, P. Georgiou, and S. Narayanan, "Automatic estimation of Parkinson's disease severity from diverse speech tasks," in *INTERSPEECH*, 2015, pp. 914–918.

[9] T. Lee, Y. Liu, Y. T. Yeung, T. K. T. Law, and K. Y. S. Lee, "Predicting severity of voice disorder from DNN-HMM acoustic posteriors," in *INTERSPEECH*, 2016, pp. 97–101.

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

[11] T. Lee, W. K. Lo, P. C. Ching, and H. Meng, "Spoken language resources for Cantonese speech processing," vol. 36, 2002, pp. 327–342.

[12] T. K. T. Law, K. Y. S. Lee, J.-H. Lam, A. C. van Hasselt, and M. C.-F. Tong, "The construction of the Cantonese perceptual evaluation of voice (CanPEV): the content validation process," in *The 4th World Voice Congress Proceedings, World Voice Consortium, Seoul, Korea*, 2010, p. 159.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," 2011.

[14] R. Fung and B. Bigi, "Automatic word segmentation for spoken Cantonese," in *Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2015 International Conference*. IEEE, 2015, pp. 196–201.