INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Predicting severity of voice disorder from DNN-HMM acoustic posteriors

Tan Lee[1,2]*, Yuanyuan Liu*[1,2]*, Yu Ting Yeung*[3]*, Thomas K.T. Law*[4]*, Kathy Y.S. Lee*[2,4]

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

Abstract

Acoustical analysis of speech is considered a favorable and promising approach to objective assessment of voice disorders. Previous research emphasized on the extraction and classification of voice quality features from sustained vowel sounds. In this paper, an investigation on voice assessment using continuous speech utterances of Cantonese is presented. A DNN-HMM based speech recognition system is trained with speech data of unimpaired voice. The recognition accuracy for pathological utterances is found to decrease significantly with the disorder severity increasing. Average acoustic posterior probabilities are computed for individual phones from the speech recognition output lattices and the DNN soft-max layer. The phone posteriors obtained for continuous speech from the mild, moderate and severe categories are highly distinctive and thus useful to the determination of voice disorder severity. A subset of Cantonese phonemes are identified to be suitable and reliable for voice assessment with continuous speech.

Index Terms: voice disorder, acoustical assessment, posterior probabilities, automatic speech recognition

1. Introduction

Use of voice is a major part of our daily life. It is not only for speech communication but also for singing, identifying a person, expressing an emotion and many other purposes. Voice problems have become very common nowadays. One of the major causes is the misuse of voice, which may be related to occupation and personal life style. Voice disorders are very common among patients with Parkinson's disease [1], and may also be caused by trauma or injury to the head and neck region [2].

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" [3]. Abnormal voices are described as being hoarse, breathy, weak, and tremorous. Currently clinical assessment of voice is carried out typically by perceptual evaluation of elicited speech samples. It aims at determining the type or severity of impairment and/or identifying specific aspects of the pathology. The accuracy and reliability of perception-based assessment depend significantly on the clinician's subjective judgement and professional experience.

Acoustical analysis of speech signals is considered a favorable and promising approach to objective assessment of voice disorders. Early studies were focused mainly on extracting feature parameters that directly quantify abnormal temporal perturbations and waveform irregularities [4][5]. Statistical modeling and pattern recognition techniques have been successfully applied to automatic classification and detection of voice pathology [6][7]. In most cases, the speech materials were limited to sustained vowel sounds. Being produced in a controlled manner, sustained vowels provide an invariant representation of steady-state phonation. Acoustic parameters extracted from sustained vowels are invulnerable to linguistic variation and other phonation-irrelevant factors, making the decision process simple and straightforward.

Many voice problems are not revealable in sustained vowels. It was shown that segmental and suprasegmental linguistic factors of connected speech, especially at consonant-vowel transitions, had strong influence on voice quality [8]. It was also found that perceptual assessment using connected and conversational speech was more reliable than sustained vowels [9]. For practical applications, the use of natural speech is preferable for its ecological validity and generalisability [7]. On the other hand, voice disorder is often investigated as a subproblem of multifaceted speech impairment, e.g., in Parkinson's disease, for which the assessment relies on a variety of speech materials [10]. When continuous speech utterances are used for voice assessment, more sophisticated methods of pattern clustering and modeling are needed to cope with the large variation of acoustic parameters. Specifically, the techniques of automatic speech recognition (ASR) can be used to pre-process input speech and facilitate feature extraction from targeted sound units. In [11], automatic rating of Parkinson's disease severity was performed using voice data from diverse speaking tasks. An ASR system was used to generate phone-level transcriptions for predicting the task type of each input utterance. In [7], the ASR forced alignment method was applied to obtain phoneme boundaries for continuous utterances of pathological speech. Voice quality parameters, including jitter, shimmer and harmonic-to-noise ratio, were computed from the vowel segments in each utterance.

This paper describes an investigation on acoustical assessment of voice disorder in Cantonese-speaking patients. In a preliminary study [12], it was noted that a Cantonese ASR system trained with normal speech exhibited significantly degraded performance for dysphonia speech. The recognition accuracy showed a strong relation with the severity level of voice disorder. This motivated us to carry out the following analysis on ASR posterior probabilities for dysphonia utterances. The hypothesis is that ASR posteriors are a good measure of the acoustic mismatch caused by voice quality change and thus can be exploited for classification and assessment of voice disorder. Another goal of this study is to identify the phonemic units that are most suitable for acoustical voice assessment.

2. Speech Database of Dysphonia Voice

The MEEI (also known as MEEI-KayPENTAX) database is by far the most commonly used database of pathological voice [13][14]. It contains the speech from 53 normal and 657 impaired subjects, each producing a sustained vowel and

http://dx.doi.org/10.21437/Interspeech.2016-1098

a continuous sentence of the same content in America English. NKI CCRT is a Dutch database of continuous sentences recorded from 55 head and neck cancer patients under chemoradiotherapy treatment [2]. This database was used in the INTERSPEECH 2012 Speaker Trait Sub-Challenge for intelligibility assessment of pathological speech [7]. Other databases of pathological voice include Saarbruecken Voice Database (SVD) in German [15] and Arabic voice pathology database (AVPD) [16]. For other languages including Chinese, voice databases of similar scale are rarely seen.

CanPEV is a Cantonese voice database developed by the Division of Speech Therapy of the Chinese University of Hong Kong (CUHK). It was designed for professional training of speech therapists on voice assessment. 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: 3 repetitions of sustained vowels /aa/, /i/ and /u/. Each repetition is about 3 to 5 second long;

Passage reading: Read-style speech of a passage that contains a brief introduction about Hong Kong. The passage consists of 146 Chinese characters and the speech is about 40 second long;

Spontaneous speech: instantaneous spoken responses to the questions "What have you done today ?" and "How do you comment on your own voice ?"

Perceptual rating was performed by 41 experienced listeners and 7 experts on voice assessment. They were asked to listen to all speech materials from each subject and to rate the voice on overall severity and a number of pre-defined vocal parameters, e.g., roughness, breathiness, strain. The ratings were given on a 10-point scale. In this study, we consider only the rating of overall severity, which is given by the average score over the 48 raters. Based on the numerical ratings, the subjects were divided into 4 categories: **normal**, **mild**, **moderate** and **severe**. In the present study, only the passage-reading utterances are used.

3. Cantonese ASR System

Cantonese is a major Chinese dialect 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 Cantonese, each Chinese character is pronounced as a monosyllable carrying a specific lexical tone. The syllable can be divided into two parts: the *Initial* (onset), and the *Final* (rime). The *Initial* is typically a consonant, while the *Final* contains a vowel nucleus followed by an optional consonant coda. 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 [17].

A Cantonese ASR system is developed to facilitate acoustical assessment of pathological voice. *Initials* and *Finals* are used as the basic units for acoustic modeling [17]. The acoustic models are trained with $15,605$ utterances from the CUSENT database and 383 passage-reading utterances (30 normal subjects) from CanPEV. CUSENT is a large-scale continuous speech database of Cantonese developed by the Chinese University of Hong Kong [18]. The speech content consists of $5,100$ distinct sentences selected from newspaper articles.

The training of acoustic models is carried out using the Kaldi speech recognition toolkit [19]. It starts with the GMM-HMM approach. A total of 297 HMMs are trained to represent position-dependent *Initials*, *Finals* and silence. Each HMM has 3 emission states. 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. With decision-tree state tying, $2,261$ probability density functions (pdf) are obtained and they are represented by $24,024$ Gaussians.

Subsequently a DNN-HMM based system is trained based on the GMM-HMM system. The acoustic feature vector is composed of 40-dimensional fMLLR features with 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 $2,261$, i.e., equal to the number of pdfs. 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.

Performance of the baseline ASR systems are evaluated with 799 test utterances from CUSENT. The pronunciation lexicon covers 630 tone-independent syllables. The language model is a syllable bi-gram trained using the orthographic transcriptions of CUSENT utterances. The syllable error rates (SER) attained by the GMM-HMM and DNN-HMM system are 10.51% and 7.82% respectively.

4. Analysis of Phone Posteriors

In this section, we first evaluate the performance of the DNN-HMM based acoustic models on pathological utterances in CanPEV. From each of the **mild**, **moderate** and **severe** categories, 10 speakers are randomly selected to contribute to the test data. The ASR performance is measured in terms of syllable error rate (SER) and phone error rate (PER), where the phone is either an *Initial* or a *Final*. Subsequently we analyze the frame-level posteriors produced by the DNN-HMM acoustic models and compare the distributions of different severity categories.

4.1. ASR Accuracy on Dysphonia Speech

Since our main focus is on the acoustic mistmatch caused by voice disorder, the effect of language model is minimized by using a uniform syllable uni-gram, i.e., the 630 syllables are assumed to be equally probable. Table 1 shows the SER and PER for the three severity categories.

Table 1: Recognition performance of DNN-HMM on pathological utterances

	Mild	Moderate	Severe
SER	6.51%	16.10%	39.35%
PER	3.08%	9.28%	24.03%

The speech recognition accuracy shows a clear trend of declining from the **mild** category to **moderate** and **severe**. This indicates an increasing degree of mismatch between the acoustic models and the pathological voices.

4.2. Phone posteriors computed from ASR output lattice

Figure 1 gives an example that explains how to obtain frame-level phone posteriors from the ASR system. It shows a speech segment that is decoded as two successive phones /h/ and /oeng/ [1]. In addition to the phone sequence, an output lattice can be obtained by using the '`lattice-to-post`" function in Kaldi. Each pair of nodes correspond to the beginning and ending points of a frame. An arc connecting the nodes is associated with a hypothesized phone and its posterior probability. If there are multiple arcs for a frame, the posteriors would sum to 1.

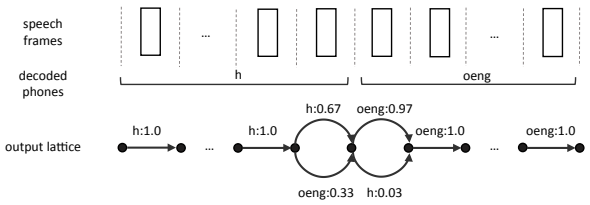

Figure 1: *Frame-level phone posteriors in ASR output lattice*

For a specific phone (*Initial* or *Final*), the average phone posterior is computed by the following procedures:

1. obtain ASR output lattices for all test utterances concerned;
2. identify the arcs that correspond to the target phone;
3. compute the average of the frame-level posteriors of the arcs.

For example, in all test utterances from the **mild** category, there are a total of 279 arcs that correspond to the *Final* /aa/. The average posterior for /aa/ of the **mild** category is computed from these arcs.

The content of CanPEV passage-reading speech covers 18 *Initials* and 35 *Finals*. The average phone posterior is computed for each of them. The distributions of these phone posteriors for the three severity classes are shown as in Figure 2. In the low posterior range (e.g., below 0.3), the number of phones from the **severe** category is much greater than that from **mild**. Whilst in the high posterior range (e.g., above 0.5), the number of phones from **severe** is the lowest and that from **mild** is the highest.

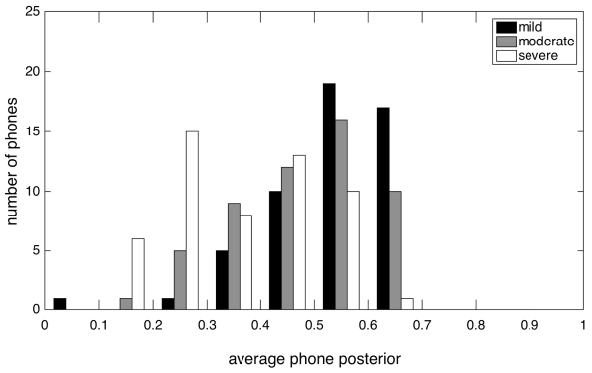

Figure 2: *Distributions of phone posteriors computed from ASR output lattice*

[1] In this paper, Cantonese *Initials* and *Finals* are represented using the Jyut Ping system[17]

4.3. Phone posteriors derived from DNN soft-max layer

Phone posteriors can also be estimated from the DNN soft-max layer as described below. At each time frame, the soft-max layer outputs a $2,261$-dimension vector of state posterior probabilities. Each of the $2,261$ state pdfs is tied to a specific phone, and each phone may be associated with multiple pdfs.

For a specific phone, the average phone posterior is computed by the following procedures:

1. perform forced alignment on all test utterances concerned;
2. identify time frames that are aligned to the target phone;
3. for each of these time frames, obtain the soft-max output vector and sum the vector elements tied to the target phone;
4. take average of the frame-level posteriors.

Similar to Section 4.2, the average phone posteriors are computed for all *Initials* and *Finals* in the CanPEV passage-reading utterances. The distributions of these phone posteriors for the three categories are shown as in Figure 3. It is seen that different severity categories are well separated by the phone posteriors. For **severe**, there are 43 phones (out of 53) having posteriors below 0.4. Whereas for **mild** and **moderate**, the number of low-posterior phones are 2 and 13, respectively.

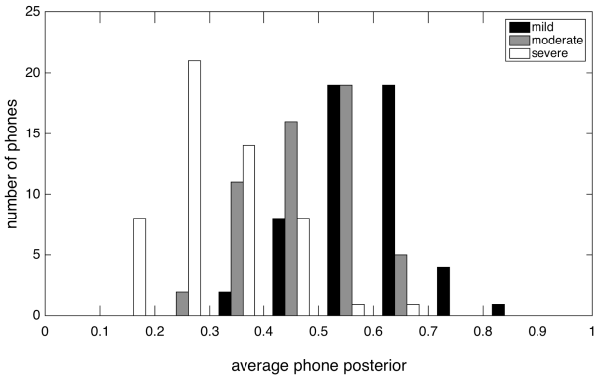

Figure 3: *Distributions of phone posteriors computed from DNN soft-max layer*

4.4. Phone matching rate

As described above, the DNN soft-max layer generates $2,261$ state posteriors, each being tied to one of the modeled *Initials* and *Finals*. A frame-level phone label can be obtained by identifying the state with the highest posterior. If the phone label matches with that given by forced alignment, this frame is said to be a matched frame. For a specific phone, the phone matching rate (PMR) is computed as follows,

$$\text{PMR} = \frac{\text{No. of matched frames}}{\text{Total no. of frames aligned to the phone}}. \quad (1)$$

For example, the test utterances from one of the subjects contain 53 frames assigned to /aai/ in the forced alignment result. From the DNN output, 49 frames have matched phone label. The value of PMR for this subject is equal to 0.92. In the subsequent analysis, the PMR is computed for the test utterances from all subjects in each category.

Similar to Section 4.2 and 4.3, the distributions of PMR for different severity categories are plotted as in Figure 4. The number of phones with PMR of 0.6 or below are 3, 7 and 41 for **mild**, **moderate** and **severe** respectively.

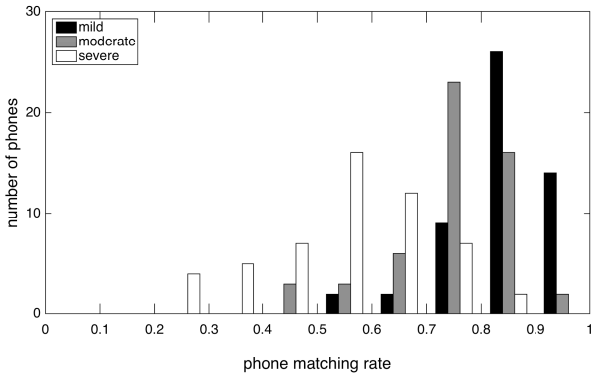

Figure 4: *Distributions of phone matching rates for the three severity categories*

5. Discussion

In the previous section, two different ways of deriving phone posteriors from the ASR system have been investigated. The analysis results on CanPEV continuous utterances clearly show that the phone posteriors have strong correlation with the categorical severity labels obtained by perceptual evaluation. The distributions of phone posteriors computed from the **severe** and the **mild** categories are very distinctive. The posterior values obtained from the **moderate** category are distributed between **severe** and **mild**.

Comparing Figure 3 and Figure 2, the phone posteriors derived from DNN soft-max layer are more discriminative in severity classification than those from ASR lattice. From Figure 4, the PMR is also expected to be more discriminate than lattice based posteriors. It must be noted that the computation of soft-max layer posteriors and PMR involves the use of additional information, i.e., the content of input speech for forced alignment. Such information may be unavailable or only partially available in real-world applications.

The plots in Figures 2, 3 and 4 do not give detailed information about individual phones. It is expected that some of the phones are more useful than the others in detecting and classifying voice disorder. There are 5 phones that have high PMR values in both **mild** and **severe**. All of them are *Initials*. On the other hand, there are 10 *Finals* that have small values of PMR in both **mild** and **severe**. This implies that *Finals* are subject to greater variation when voice disorder is present. Each *Final* contains a vowel nucleus, which could be /aa/, /e/, /i/, /o/, /u/, /yu/ or /oe/. We group the *Finals* that has the same vowel nucleus, and compute the PMR for the respective broad class of *Finals*. As shown in Figure 5, the PMR values of **mild** and **severe** are highly contrastive, while the **moderate** category takes values in-between.

Let P_{mild} and P_{severe} denote the soft-max posteriors of a specific phone in the **mild** and the **severe** categories respectively. The ratio of P_{mild} to P_{severe} reflects how effective this phone is in distinguishing **severe** disorder from **mild**. For each of the 18 *Initials* and 53 *Finals*, the posterior ratio is computed. In a similar manner, the PMR ratio between **mild** and **severe** is

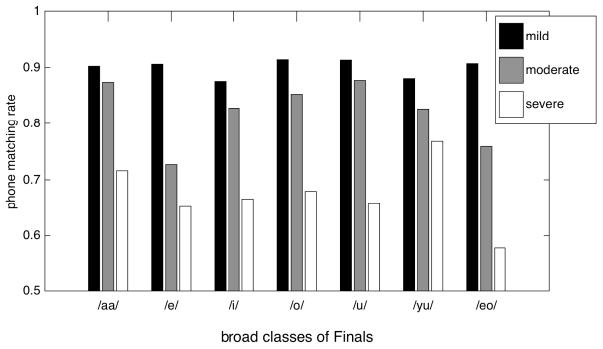

Figure 5: *Phone matching rates of broad classes of Finals*

obtained. By examining the posterior ratios and PMR ratios of all *Initials* and *Finals*, the following observations can be made:

1. The posterior ratio and the PMR ratio for the same phone are highly correlated;
2. Phones that have large posterior ratios and PMR ratios are mostly *Finals*. The 10 most discriminative *Finals* are: /aa/, /ak/, /ang/, /un/, /aak/, /aam/, /ui/, /u/, /e/ and /aai/;
3. Some of the voiced *Initials*, namely, /ng/, /n/ and /l/, have large posterior ratios that are comparable to the most discriminative *Finals*;
4. The unvoiced *Initials*, namely /p/, /z/, /f/, /c/ and /s/, are least discriminative in detecting voice disorder. They have posterior ratios and PMR ratios close to 1.

6. Conclusions

Using continuous speech for objective assessment of voice disorder is practically feasible and clinically appealing. The present study shows that phone posteriors produced by a DNN-HMM ASR system are effective in quantifying and predicting the severity of voice disorder. Although the research has been based on Cantonese speech, we believe that the methodology and key findings are generalizable to other languages. Toward the design of an automatic assessment system, we propose to use phone-specific posteriors as input features for the detection and classification of voice disorders. For continuous speech utterances, the ASR system can be used to generate phone alignments and dysphonia features can be extracted from a specific set of discriminative phones.

7. Acknowledgements

This research is partially supported by a GRF project grant (Ref: 14204014) from Hong Kong Research Grants Council, Major Program of National Social Science Fund of China (Ref: 13&ZD189), and by the Shenzhen Municipal Engineering Laboratory of Speech Rehabilitation Technology. The CanPEV database was developed with the support of a GRF project (Ref: 468708).

8. References

[1] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[2] L. van der Molen, M. A. van Rossum, A. H. Ackerstaff, L. E. Smeele, C. R. Rasch, and F. J. Hilgers, "Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views," *BMC Ear, Nose and Throat Disorders*, vol. 9, no. 1, p. 1, 2009.

[3] S. Taylor-Goh, "RCSLT Clinical Guidelines," 2005.

[4] E. H. Buder, "Acoustic analysis of voice quality: A tabulation of algorithms 1902–1990," *Voice quality measurement*, pp. 119–244, 2000.

[5] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1264–1271, 2012.

[6] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 10, pp. 1943–1953, 2006.

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

[8] 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.

[9] T. Law, J. H. Kim, K. Lee, E. Tang, J. Lam, A. C. van Hasselt, and M. C. Tong, "Comparison of raters reliability on perceptual evaluation of different types of voice sample," *Journal of Voice*, vol. 26, no. 5, 2012.

[10] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The Interspeech 2015 computational paralinguistics challenge: nativeness, Parkinsons & eating condition," in *Proceedings of Interspeech*, 2015.

[11] 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 parkinsons disease severity from diverse speech tasks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 914–918.

[12] T. Lee, Y. Liu, P.-W. Huang, J.-T. Chien, W.-K. Lam, Y.-T. Yeung, T. Law, K. Lee, A. P.-H. Kong, and S.-P. Law, "Automatic speech recognition for acoustical analysis and assessment of Cantonese pathological voice and speech," in *Proceedings of ICASSP*, 2016, pp. 6475–6479.

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

[14] "Disordered voice database and program, model 4337," http://www.kayelemetrics.com/, KayPENTAX, NJ.

[15] W. J. Barry and Pützer, "Saarbruecken voice database," http://www.stimmdatenbank.coli.uni-saarland.de/, Institute of Phonetics, University of Saarland.

[16] A. Al-nasheri, Z. Ali, G. Muhammad, and M. Alsulaiman, "An investigation of mdvp parameters for voice pathology detection on three different databases," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 2952–2956.

[17] 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.

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

[19] 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.