Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 119–134

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams [☆]

Xing Fan, John H.L. Hansen [*]

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA

Received 8 August 2011; received in revised form 6 July 2012; accepted 24 July 2012
Available online 25 August 2012

Abstract

Whispered speech is an alternative speech production mode from neutral speech, which is used by talkers intentionally in natural conversational scenarios to protect privacy and to avoid certain content from being overheard or made public. Due to the profound differences between whispered and neutral speech in vocal excitation and vocal tract function, the performance of automatic speaker identification systems trained with neutral speech degrades significantly. In order to better understand these differences and to further develop efficient model adaptation and feature compensation methods, this study first analyzes the speaker and phoneme dependency of these differences by a maximum likelihood transformation estimation from neutral speech towards whispered speech. Based on analysis results, this study then considers a feature transformation method in the training phase that leads to a more robust speaker model for speaker ID on whispered speech without using whispered adaptation data from test speakers. Three estimation methods that model the transformation from neutral to whispered speech are applied, including convolutional transformation (ConvTran), constrained maximum likelihood linear regression (CMLLR), and factor analysis (FA). a speech mode independent (SMI) universal background model (UBM) is trained using collected real neutral features and transformed pseudo-whisper features generated with the estimated transformation. Text-independent closed set speaker ID results using the UT-VocalEffort II corpus show performance improvement by using the proposed training framework. The best performance of 88.87% is achieved by using the ConvTran model, which represents a relative improvement of 46.26% compared to the 79.29% accuracy of the GMM-UBM baseline system. This result suggests that synthesizing pseudo-whispered speaker and background training data with the ConvTran model results in improved speaker ID robustness to whispered speech.

Keywords: Speaker identification; Whispered speech; Vocal effort; Robust speaker verification

1. Introduction

Whispered speech is a natural speech production mode, employed in public situations in order to protect privacy and to avoid certain content from being made public. For example, a customer might whisper to provide information regarding their date of birth, credit card information, and billing address in order to make hotel, flight, or car reservations through a machine interface over the telephone, or a doctor might whisper when entering a voice memo in order to discuss patient medical records in public. Aphonic individuals, as well as those with low vocal

[☆] This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067 (Approved for public release, distribution unlimited), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

[*] Corresponding author. Address: Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Dept. of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA. Tel.: +1 972 883 2910; fax: +1 972 883 2710.

E-mail address: John.Hansen@utdallas.edu (J.H.L. Hansen).

URL: http://crss.utdallas.edu (J.H.L. Hansen).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.07.002

capability, such as heavy smokers, also employ whisper as a primary form of oral communication. In this study, the term "neutral speech" refers to speech produced at rest in a quiet sound-booth whose "voiced" phonemes, such as sustained vowels, contain glottal based vocal fold movement that represents periodic excitation.

There are significant differences between whisper and neutral speech production mechanisms, which result in the absence of voiced excitation, shifted formant locations and change in formant band width (Ito et al., 2005; Zhang and Hansen, 2007; Morris and Clements, 2002; Matsuda and Kasuya, 1999; Jovicic, 1998). Zhang and Hansen (2007) revealed that the change of vocal effort in test data ranging from whisper through shouted has a significant impact on automatic speaker identification (speaker ID) performance, with whisper resulting in the most serious loss in performance. Similar results were reported in other studies on automatic speech recognition (Ito et al., 2005) and speaker recognition (Jin et al., 2007) systems as well.

Past work on automatic speaker ID systems for whispered speech can be grouped into two main categories: front-end processing (Fan and Hansen, 2009; Fan and Hansen, 2008) and back-end model adaptation (Jin et al., 2007). Both methods have resulted in improvements in system accuracy. However, new front-end processing methods involve feature re-extraction and model re-training for neutral speech, which increases computational requirements and may hurt system performance on neutral test speech. For back-end model adaptation, as in Jin et al. (2007), a simple maximum a posteriori (MAP) adaptation of the original model trained with neutral speech can provide satisfactory performance under the prerequisite of a fair amount of speaker-dependent (SD) whispered adaptation data. However, in real applications, whispered adaptation data from test speakers is generally not available. Also, while it is possible to collect additional whispered data from other speakers, the fact that the total amount of real whispered data is usually much smaller compared with the available neutral data means that it is still very difficult to train a speech mode independent (SMI) universal background model (UBM). Therefore, the focus of this study is to explore efficient model training techniques that rely solely on a limited set of whisper data from *non-target speakers* for modeling whispered speech. In this study, *non-target speakers* are those speakers whose speech is not seen in the test set for closed-set speaker ID.

A similar strategy was first considered by Bou-Ghazale and Hansen (1998), where HMMs were used to statistically model characteristics needed for generating pitch contour and spectral slope patterns in order to modify the speaking style from neutral to stressed speech. In this study, the statistical information contained in a UBM trained with whispered data set collected from *non-target speakers* is employed for a transformation estimation to generate whisper features from neutral data. The convectional Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980), which are employed for most state-of-the-art speech systems, are used here as the front-end features throughout this study and our compensation is applied in the corresponding MFCC domain. The generated whispered features will be referred as "pseudo-whisper features" in the rest of this study.

Formulating a model training method for this task requires understanding two critical facets of the problem. One is the difference between whispered and neutral speech in the resulting front-end feature domain. In particular, the MFCCs represent information regarding the smoothed spectral envelope in the Mel domain, hence, the differences between whisper and neutral in the linear frequency domain (Ito et al., 2005; Morris and Clements, 2002), might be distorted and represented in a different way in the Mel domain. The other facet is the consistency of the differences among speakers and phonemes. For example, if the spectral differences between whispered and neutral speech are consistent across speakers, a transformation estimated using whispered adaptation data from several *non-target speakers* could be applied directly to all whispered enrollment and test data for automatic speaker ID. On the other hand, if spectral differences between whispered and neutral speech are inconsistent across speakers (i.e., the way someone "whisper" may be speaker dependent), it is necessary to explore alternative methods that could estimate the particulars of a given enrollment speaker's whispered speech. If the spectral differences are phoneme or phoneme-class dependent, the problem will be even more complex since a unique mapping will be needed for each phoneme or phoneme-class. Past studies (Ito et al., 2005; Jovicic, 1998; Matsuda and Kasuya, 1999; Eklund and Traunmuller, 1996) provided comparison results for the average differences between whispered and neutral speech across phonemes in the linear frequency domain. However, those studies have not examined individual speaker differences in terms of the variations of those differences in the linear frequency or the Mel domain.

This study first compares the smoothed spectral envelope of whispered and neutral speech using a maximum likelihood transformation estimation. The dependence of the estimated transformation on speakers and phonemes is analyzed. Based on the analysis results, this study proposes a method that models the differences between whispered and neutral speech by a convolutional filter with zero mean additive noise (ConvTran). The parameters of the ConvTran transformation are estimated using a first order vector Taylor series (VTS) approximation and the expectation maximization (EM) algorithm. Pseudo-whisper features generated with the proposed ConvTran model are used to train a SMI-UBM, which will include equal amounts of neutral and pseudo-whispered speech. Also, because the proposed method keeps some level of speaker-dependent information in the resulting pseudo-whisper features, after the SMI-UBM is trained, a speaker dependent model can be further obtained by adaptation of the SMI-UBM with both neutral and selected pseudo-whisper features. Constrained maximum likelihood linear regression (CMLLR) and factor analysis (FA) transformation models are also applied for the purpose of performance comparison.

The remainder of this paper is organized as follows: Section 2 introduces the production and acoustic characteristics of whispered speech. Section 3 describes the constructed corpus employed for acoustic analysis and speaker ID. Section 4 introduces the transformation estimation method based on VTS and EM algorithms. Section 5 discusses the analysis methods and results. Section 6 describes the speaker ID system and subsequent experimental results. Finally, Section 7 discusses the conclusions.

2. Whispered speech

In neutral speech, voiced phonemes are produced through a periodic vibration of the vocal folds to produce glottal air flow into the pharynx, oral cavities and nasal cavities. However, for whispered speech, the vocal folds remain open without vibration, resulting in a continuous uninterrupted air stream with no periodic excitation. The air flow from the lungs is used as the excitation sound source, and the shape of the pharynx is adjusted such that the vocal folds will not vibrate (Thomas, 1969; Gavidia-Ceballos and Hansen, 1996; Meyer-Eppler, 1957). In order to illustrate this, Fig. 1 shows the significant differences in waveform and spectrogram characteristics of the speech signal “Don’t do Charlie’s dirty dishes” from the same speaker between neutral and whisper modes. Clearly, the time waveform for whisper speech is significantly lower in amplitude and the complete absence of voiced excitation is obvious in both the whispered speech time waveform and spectrogram. Other variations, such as duration differences, can be observed as well.

The differences in the speech production process between whispered and neutral speech are reflected in the following aspects in the spectral domain: first, there is no periodic excitation or harmonic structure in whisper. Second, the location of lower frequency formants in whispered speech is generally shifted to higher frequencies compared to neutral speech (Ito et al., 2005). Third, the spectral slope of whispered speech is much flatter than that of neutral speech, and the duration of whispered speech is longer than that of neutral speech (Zhang and Hansen, 2007). Fourth, the boundaries of vowel regions in the F1–F2 frequency space also differ from neutral speech (Eklund and Traunmuller, 1996; Kallail et al., 1984). Finally, whispered speech has a much lower energy contour compared with the same neutral speech sequence. Due to these differences, traditional neutral speech trained speaker ID systems degrade significantly when tested with whispered speech. From a physiological perspective, it is possible that an equivalent “amount” of speaker dependent information as that seen in neutral speech is present, but the resulting speech features and speaker model are not capable of characterizing this content. Alternatively, it is possible that the speaker dependent information is actually lost or not conveyed under whispered speech. If this is the case, it may not be theoretically possible to achieve the same level of speaker ID performance for whispered speech as that seen for neutral speech.

3. Corpus description

In order to confirm the validity of the analysis results as well as the effectiveness of the proposed system, this study employs two corpora: UT-VocalEffort I and II for acoustic analysis and speaker ID system development, respectively.

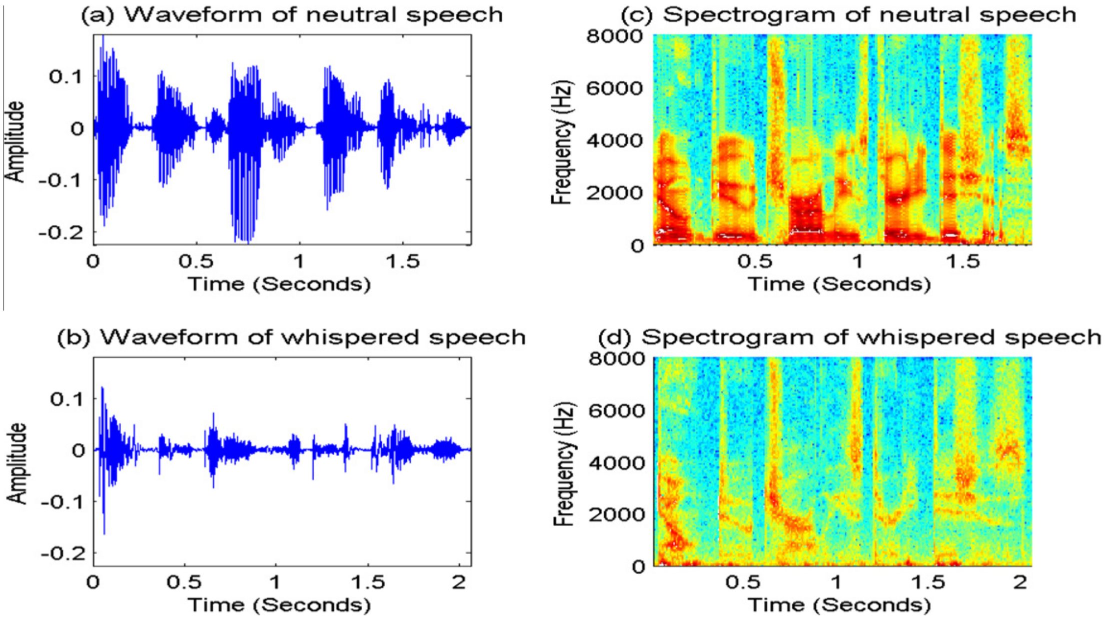

Fig. 1. Time domain waveforms and spectrogram of the speech signal “Don’t do Charlie’s dirty dishes” from the same speaker in (a,c) neutral and (b,d) whispered speech mode.

The UT-VocalEffort I corpus (Zhang and Hansen, 2007) supplies the whispered/neutral paired utterances used in the acoustic analysis of this study. Ten male native speakers of American English were recruited to speak ten sentences drawn from the TIMIT database in both whisper and neutral modes (Garofolo et al., 1993). This corpus has the advantage that for phone level acoustic analysis, the same phoneme context is provided across both speakers as well as speech modes, and thus any dependency resulting from different phoneme distributions is minimized.

For the development and evaluation of speaker ID system for whispered speech, the UT-VocalEffort II corpus developed in Zhang and Hansen (2009) is employed. This corpus consists of a total of 112 speakers, 37 males and 75 females. Whispered and neutral speech from 28 native-speaking American English female subjects are chosen for the development of a closed-set speaker ID recognition system. For simplicity, this set is referred as the NW28 set for the remainder of this paper. In the NW28 set, each speaker has an average 4.5 minutes neutral training data and an average of 34 whispered test utterances ranging from 1 to 3 s. The neutral speech from NW28 set is used for GMM training, and the whispered speech from NW28 is used as test data for automatic speaker ID. Another 10 separate female speakers' (without overlap with NW28) whispered speech are chosen as the development set and is referred as the WH10 set. The UT-VocalEffort II corpus consists of both read and spontaneous parts. In this study, only the read part is considered. In the WH10 development set, each speaker has an average of 1 min whispered data. All utterances are entire utterances from the UTVocal effort corpus instead of being segmented.

The whispered and neutral streams of all subjects were manually separated to constitute the whisper and neutral corpora. All recordings were obtained in an ASHA-certified single-walled soundbooth, using a Shure Beta 53 head-worn close talking microphone, and were digitized and recorded using a Fostex 8 D824 channel synchronized digital recorder at 44.1 kHz, with 16 bits per sample, and down sampled to 16 kHz for this study. From Zhang and Hansen (2009) and Zhang and Hansen (2007), we also note that all recordings include a 1 kHz 75 dB pure-tone calibration test sequence to provide ground-truth on true vocal effort for all speakers and sessions.

4. VTS based adaptation formulation

The VTS approximation based acoustic-model/feature adaptation is previously proposed for joint compensation of additive and convolutive distortions in robust speech recognition systems (Moreno et al., 1996; Deng et al., 2004; Li et al., 2009). The strategy is applied here with modification for the purpose of feature transformation estimation. Before it is applied in our study for acoustic analysis and automatic speaker ID in Sections 5 and 6, this section will present a general description of the speech transformation model, and the derivation for adaptation formulas of Gaussian mixture models used in our VTS based adaptation algorithm.

4.1. Speech transformation model

In the VTS expansion adaptation algorithm, the target speech $y(t)$ is assumed to be generated from the source speech $x(t)$ with a channel filter $h(t)$ and noise $n(t)$ according to:

$$y(t) = x(t) * h(t) + n(t). \tag{1}$$

For simplicity, we assume the cosine of the angle between $x(t) * h(t)$ and $n(t)$ in frequency domain equals zero. Thus, in the MFCC domain, the relationship between $y(t)$ and $x(t)$ can be represented as:

$$y = x + h + g(x, h, n), \tag{2}$$

$$g(x, h, n) = \mathrm{C}\log(1 + \exp(C^{-1}(n - x - h))), \tag{3}$$

where C^{-1} is the pseudo-inverse DCT matrix and y, x, h, n are the MFCCs for $y(t), x(t), h(t)$ and $n(t)$ respectively. In this study, the noise n is assumed to be Gaussian distributed with zero mean μ_n and a diagonal covariance matrix Σ_n. The channel filter h is assumed to be a fixed vector with deterministic values that represents the shape of the smoothed spectral envelope of $h(t)$. Assuming μ_x is the mean of x and applying the first order VTS approximation around the point (μ_x, μ_h, μ_n), we have

$$y \approx \mu_x + \mu_h + g(\mu_x, \mu_h, \mu_n) + G(x - \mu_x) + G(h - \mu_h) \\ + F(n - \mu_n), \tag{4}$$

where,

$$\begin{aligned} &\frac{\partial y}{\partial x}\Big|_{\mu_x,\mu_h,\mu_n} = \frac{\partial y}{\partial h}\Big|_{\mu_x,\mu_h,\mu_n} = G \\ &\frac{\partial y}{\partial n}\Big|_{\mu_x,\mu_h,\mu_n} = I - G = F \\ &G = C \cdot diag\left\{\frac{1}{1+\exp(C^{-1}(\mu_n - \mu_x - \mu_h))}\right\} \cdot C^{-1}, \end{aligned} \tag{5}$$

where $diag\{\}$ stands for a diagonal matrix with its diagonal component value equal to the value of the vector in the argument. Taking the expectation and variance operations of both sides of Eq. (4), the resulting static μ_y and Σ_y are (noting that the filter h is a fixed vector):

$$\begin{aligned} &\mu_y \approx \mu_x + \mu_h + g(\mu_x, \mu_h, \mu_n), \\ &\Sigma_y \approx G\Sigma_x G^t + F\Sigma_n F^t. \end{aligned} \tag{6}$$

For the dynamic feature vectors (delta and delta/delta portions of MFCC features), the following hold Li et al. (2009):

$$\begin{aligned} &\mu_{\Delta y} \approx G\mu_{\Delta x} \\ &\mu_{\Delta\Delta y} \approx G\mu_{\Delta\Delta x} \\ &\Sigma_{\Delta y} \approx G\Sigma_{\Delta x} G^t + F\Sigma_n F^t \\ &\Sigma_{\Delta\Delta y} \approx G\Sigma_{\Delta\Delta x} G^t + F\Sigma_n F^t. \end{aligned} \tag{7}$$

Given that μ_x^{sm} is the mean of the mth Gaussian and Σ_x^{sm} is the covariance matrix of the mth Gaussian in the sth state from source speech x's models (either GMMs or HMMs), Eq. (6) and Eq. (7) can be employed for updating the corresponding Gaussian pdf parameters.

4.2. EM algorithm for estimation of parameters

The EM algorithm (Dempster et al., 1977; Li et al., 2009) is applied to estimate μ_h. Given a target utterance y, the EM auxiliary function is:

$$Q(\lambda|\bar{\lambda}) = \sum_t \sum_{s,m} \gamma_{tsm} \log(p(y_t|s,m,\lambda)), \tag{8}$$

where $p(y_t|s,m,\lambda) \sim N(y_t; \mu_{x,sm}, \Sigma_{x,sm})$ and γ_{tsm} is the posterior probability of the mth Gaussian pdf in the sth state of HMM for the tth frame in y.

In the M-step, we take the derivatives of Q with respect to μ_h. The update formula for each μ_h is found by setting the derivatives to zero:

$$\mu_h = \mu_{h,0} + \left\{ \sum_t \sum_{s,m} \gamma_{tsm} G^t_{s,m} \Sigma^{-1}_{x,sm} G_{s,m} \right\}^{-1} \times \left\{ \sum_t \sum_{s,m} \gamma_{tsm} G^t_{s,m} \Sigma^{-1}_{x,sm} [y_t - \mu_{x,sm} - \mu_{h,0} - g(\mu_{x,sm}, \mu_{x,sm}, \mu_n)] \right\}. \tag{9}$$

The noise term n is assumed to be short-time stationary, thus $\mu_{\Delta n} = 0$ and $\mu_{\Delta\Delta n} = 0$. The Σ_n is updated as in Li et al. (2009) using Newton's method:

$$\Sigma_n = \Sigma_{n,0} - \left[\left(\frac{\partial^2 Q}{\partial^2 \Sigma_n} \right)^{-1} \left(\frac{\partial Q}{\partial \Sigma_n} \right) \right]. \tag{10}$$

For $\Sigma_{\Delta n}$ and $\Sigma_{\Delta\Delta n}$, a similarly derived update formula is employed. The next two sections will consider the application of VTS adaptation in detail.

5. Acoustic analysis for phoneme and speaker dependency

This section explores how the differences between whispered and neutral speech depend on speakers and phonemes. In particular, this section models the transformation of neutral speech $y_{\mathbf{ne}(t)}$ towards whisper $x_{\mathbf{wh}(t)}$ using a linear time-invariant (LTI) filter h(t) plus a noise term n(t) in the MFCC domain. The parameters of h(t) in the cepstral domain are obtained using a first order VTS approximation and the EM algorithm as described in Section 4. This assumption serves as a first order approximation; therefore the model is limited in its power to capture detailed spectral structure change, but will capture overall aspects of the differences between whispered and neutral speech in smoothed spectral envelope. Also, due to the introduction of convolution, the complexity of the model to be estimated decreases significantly compared with other cepstral domain linear regression, thus reducing the chance of overfitting and resulting in an estimate close to ground truth.

5.1. Experimental method

Let $y_{\mathbf{wh}(t)}$ represent the target whispered MFCC feature and $x_{\mathbf{ne}(t)}$ represent the source neutral MFCC feature. Next, the speech transformation model described in Section 4.1 can be represented as:

$$y_{\mathbf{wh}} = x_{\mathbf{ne}} + h + g(x_{\mathbf{ne}}, h, n), \tag{11}$$

$$g(x_{\mathbf{ne}}, h, n) = C\ \log(1 + \exp(\mathrm{C}^{-1}(n - x_{\mathbf{ne}} - h))). \tag{12}$$

In order to estimate h for separate phonemes and speakers, the general VTS adaptation algorithm, described in Section 4, is implemented as shown in Fig. 2. Note that we assume h is deterministic, so μ_h simply equals to h

In this section, all speech is windowed with a Hamming window of length 25 ms, with a 10 ms overlap. 13-dimensional MFCCs, appended with their first- and second-order time derivatives are used as acoustic features. Each HMM is left-to-right with 3 states and 16 Gaussian mixtures per state. In order to obtain reliable neutral HMMs used in the first block in Fig. 2, the TIMIT corpus (Garofolo et al., 1993) is employed to obtain an initial sets of HMMs. Next, using the 10 neutral sentences from each speaker in UT-VocalEffort I, speaker dependent neutral HMMs are

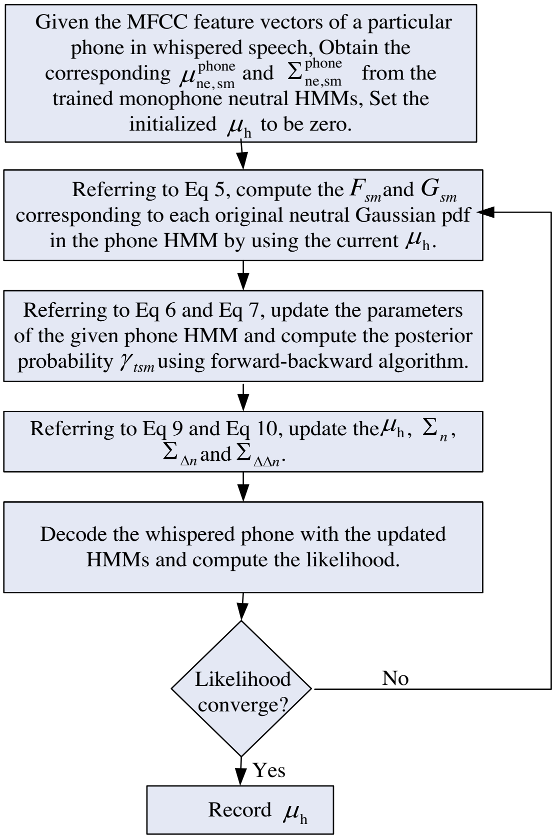
Given the MFCC feature vectors of a particular phone in whispered speech, Obtain the corresponding $\mu^{\text{phone}}_{\text{ne,sm}}$ and $\Sigma^{\text{phone}}_{\text{ne,sm}}$ from the trained monophone neutral HMMs, Set the initialized μ_{h} to be zero.
Referring to Eq 5, compute the $F_{sm}$and G_{sm} corresponding to each original neutral Gaussian pdf in the phone HMM by using the current μ_{h}.
Referring to Eq 6 and Eq 7, update the parameters of the given phone HMM and compute the posterior probability γ_{tsm}using forward-backward algorithm.
Referring to Eq 9 and Eq 10, update theμ_{h}, Σ_n, $\Sigma_{\Delta n}$and $\Sigma_{\Delta\Delta n}$.
Decode the whispered phone with the updated HMMs and compute the likelihood.
Likelihood converge?
No
Yes
Record μ_{h}

Fig. 2. VTS-based adaptation algorithm used in acoustic analysis.

adapted from the TIMIT HMMs using maximum likelihood linear regression (MLLR).

From the obtained speaker dependent neutral HMMs, $\mu^{phone}_{\mathbf{ne,sm}}$ and $\Sigma^{phone}_{\mathbf{ne,sm}}$ in Fig. 2 are simply the mean and covariance of the mth mixture Gaussian in the sth state. In order to segment all whispered speech from UT-VocalEffort I into phone level units, we adapt the TIMIT HMMs into speaker-dependent whisper HMMs using the same whispered data with transcription, and apply forced-alignment to detect the phone boundaries. Next, given each whisper phone from a particular speaker, we follow the steps in Fig. 2 and estimate h from Eq. (11) using the neutral HMM from the corresponding speaker/phone. To make sure the speaker dependent neutral HMMs that provide $\mu^{phone}_{\mathbf{ne,sm}}$ and $\Sigma^{phone}_{\mathbf{ne,sm}}$ capture enough speaker information, only phonemes with sufficient neutral adaptation data for MLLR are considered, which results in a total of 32 monophones in our study.

5.2. Analysis results

After we estimate the parameter h for each whispered phone occurrence in the UT-VocalEffort I corpus, it is converted to the frequency domain by applying C^{-1}, the pseudo inverse of the DCT matrix, and is denoted as μ_H. As an example of the resulting estimate, the left column of Fig. 3 shows the means of μ_H for the whispered phoneme /ih/, /n/, and /z/ for each speaker. To confirm the reliability of our estimation methods and its implementation, we also estimate the means of μ_H per speaker using the neutral phonemes instead of whisper phonemes. The neutral phoneme units used to obtain the right column of Fig. 3 are without overlapping with the neutral adaptation data used for obtaining speaker dependent neutral HMMs. The transformations estimated on the right column are all near zero, which indicates that the given neutral phoneme units are similar to the data used to train the corresponding neutral monophone HMMs. The left column of Fig. 3 shows that for vowel /ih/ and nasal /n/, the transformation of neutral speech into whisper includes compression of the energy of the neutral speech especially below 2 kHz. For the voiced fricative "z", the transform is near zero, implying that the neutral speech and the whispered speech are very similar in the MFCC domain. The remaining sections will divide the μ_h into vowels and consonants for separate analysis.

5.2.1. Results of the vowel analysis

Fisher's discriminant power is used to analyze the dependence of inter-speaker variation and inter-phone variation. A greater magnitude of the discriminant power implies better separation between the given clusters in the sample space. Therefore, by comparing the Fisher's discriminant power relatively under different classification criterions we can explore the dependency of the differences between whispered and neutral speech on phonemes and speakers.

Given K classes of μ_h that constitute the sample space Φ_K, there are K cluster means $\mu_{h,k}$ and K cluster diagonal covariances matrix $\Sigma_{h,k}$, where $1 \leqslant k \leqslant K$. The mean of the cluster means $\mu_{h,k}$ is denoted $\bar{\mu_g}$. Assuming there are W_k samples in each class of $\mu_{h,k}$, Fisher's discrimination power is defined as:

$$F(\mu_h) = \frac{\sum_{k=1}^{K} W_k(\mu_{h,k} - \bar{\mu_g})^T(\mu_{h,k} - \bar{\mu_g})}{\mathrm{trace}\left(\sum_{k=1}^{K} \Sigma_{h,k}\right)}. \tag{13}$$

The Fisher discrimination power F_p is computed for all speakers s, treating each phoneme p as a class. This measures the inter-phoneme variability of μ_h among all the

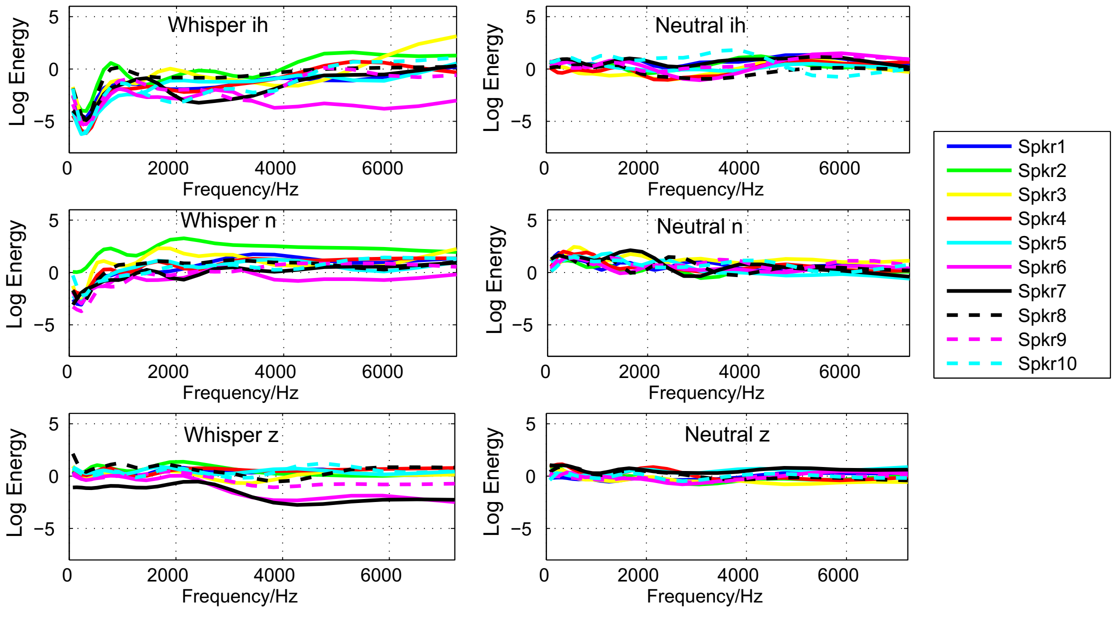

Fig. 3. Examples of the average estimated μ_H.

speakers. The Fisher discrimination power F_s is computed for all the phonemes, treating each speaker s as a class. This measures the inter-speaker variability of μ_h among all the phonemes. The F_p and F_s is calculated and listed in Table 1. A larger F_p indicates a better separation of μ_H among phonemes class, while a larger F_s indicates a better separation of μ_H among speakers.

To provide more detailed information, the frequency range from 0 to 8 kHz is arbitrarily divided into 3 subbands: S1(0–2700 Hz), S2(2700–4000 Hz), and S3(4000–8000 Hz), representing approximately a phone dependent frequency range, a speaker dependent frequency range, and the remaining high frequency range. The subbands μ_h^{S1}, μ_h^{S2}, and μ_h^{S3} can be obtained from μ_h through simple linear algebra. Fisher's discriminant power is also used to separately analyze the dependence of each subband on the inter-speaker variation F_s^{Sx} and inter-phone variation F_p^{Sx} with results listed in Table 1.

Table 1 shows that the F_s^{Sx} increases with increasing frequency subband, while in comparison F_p^{Sx} remains stable with relatively small values. This suggests that the differences between whispered vowels and neutral vowels are similar across the frequency range given a specific speaker. At lower frequencies, F_p^{S1} and F_s^{S1} shares similar values, which indicates the differences are somewhat phoneme dependent and somewhat speaker dependent. With an increase in frequency, the phoneme-dependency is lost while the speaker-dependency difference strongly remains and increases significantly.

Table 1
Fisher's discrimination power in discriminating vowels and discriminating speakers, for fullband *S* and subbands *S*1, *S*2 and *S*3.

Subband	F_p^{Sx}	F_s^{Sx}
S (fullband)	71.0	168.6
*S*1	82.7	85.3
*S*2	47.1	206.8
*S*3	48.0	447.3

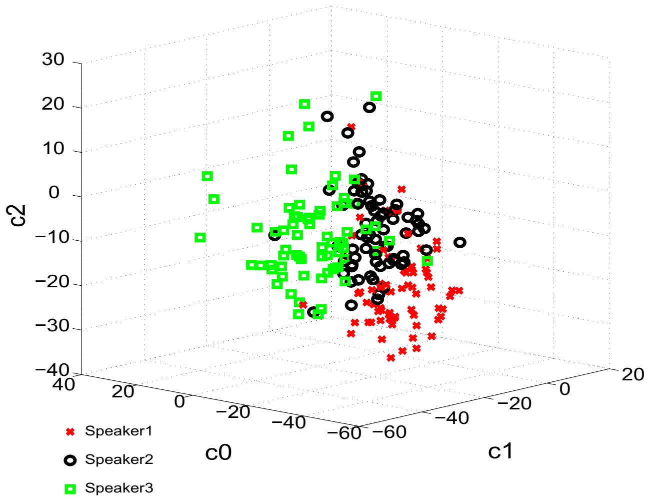

Fig. 4. Examples of c0 (energy), c1 (first order of MFCC) and c2 (second order of MFCC) for three speakers under whisper speech model using same phoneme content.

Fig. 4 confirms the above observation, where the first three orders of MFCC: c0, c1 and c2 of all μ_h estimated from vowels for three distinct speakers are plotted respectively. The three colors represent each of the three speakers and each color covers all seen vowel phonemes from the corresponding speaker. This figure shows that μ_h is clearly more speaker dependent than phoneme dependent because of the obvious separation of μ_h among the three speakers. Fig. 5 also confirms the same conclusion, where the average μ_h for each speaker in the corpus for the vowels /ax/, /ih/ and /uw/ are plotted. The figure shows that the μ_h is generally speaker dependent, especially in higher frequencies, while there are slight variations in the lower frequencies. The observations in Fig. 5 confirm results of the statistical analysis in the fullband and subband with Fisher's discrimination power.

5.2.2. Results from consonant analysis

For the following analysis, all consonants are grouped into five categories due to their different production mechanism: (1) Unvoiced consonants (UVC), which include unvoiced stops, affricates, and fricatives, (2) Voiced consonants from stops, affricates, and fricatives (VC) that can be mapped to the unvoiced consonants, (3) Nasals, (4) Liquids and (5) Glides. In order to analyze the impact of the absence of voiced excitation on consonants, the spectral tilt of the neutral-whisper transfer function μ_H is measured using a first order linear regression of $\mu_H = \mathrm{a}[\log \mathit{frequency}] + \mathrm{b}$, where a represents the spectral tilt of the estimated μ_H. Fig. 6 shows that a and b are around zero for UVC and VC, which suggests that UVC and VC share a similar spectral tilt change in the smoothed spectral envelope, despite the absence of voiced excitation in whispered VC. However, we can observe from Fig. 6 that the values of a and b for nasals, glides and liquids are much larger, which suggests that the spectral tilt of nasals, glides, and liquids undergoes a greater change from neutral to whisper than that seen for UVC and VC.

In order to investigate the speaker and phoneme dependency of μ_h for consonants, $F_s(\mu_h^{S,S1,S2,S3})$ and $F_p(\mu_h^{S,S1,S2,S3})$ are calculated in the same way as that used for the analysis of the vowels. Considering the similarity between whispered speech and neutral speech for the production of stops, fricatives and affricates, only liquids, glides, and nasals are considered in this part of the experiment. Table 2 lists the Fisher discrimination power, which shows that μ_h is highly phoneme dependent in the lower frequency range. This result is consistent with a and b in Fig. 6. Also, μ_h becomes more speaker dependent with increase in frequency band, similar to that seen for vowels.

Results from this analysis on the whisper/neutral difference dependency for phonemes and speakers suggest that for vowels, the difference between whispered and neutral speech is generally more consistent across speakers than

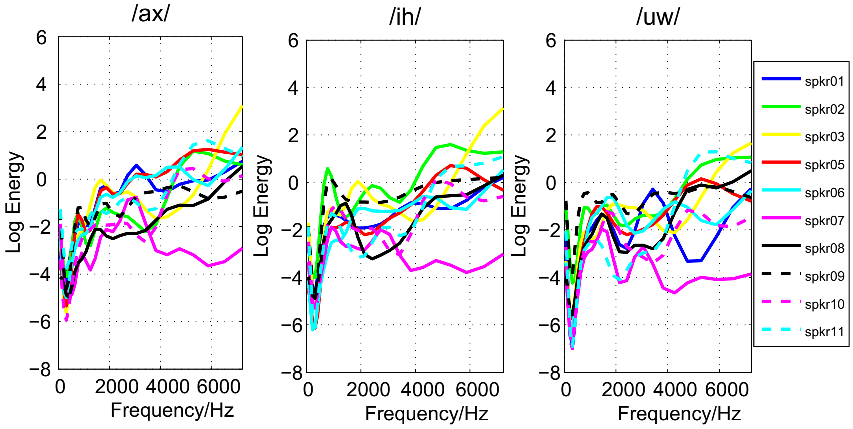

Fig. 5. Average μ_h for each speaker for each of the vowels /ax/, /ih/ and /uw/.

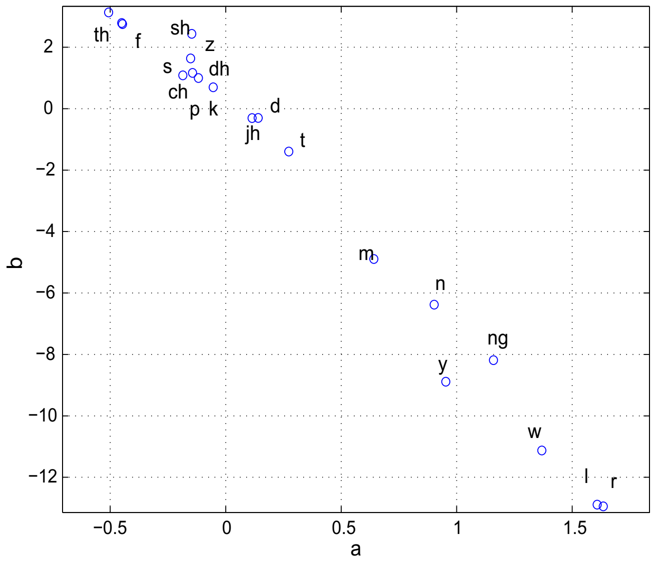

Fig. 6. Distribution of a and b for consonants.

Table 2
Fisher's discrimination power in discriminating consonants, and discriminating speakers, for fullband S, and subband S1, S2 and S3.

Subband	F_p^{Sx}	F_s^{Sx}
S (full band)	151.2	60.8
S1	210.5	33.5
S2	83.6	70.5
S3	76.2	137.2

phonemes, especially beyond 3 kHz. The results also suggest that the differences between whispered and neutral speech are somehow less pattern tractable below 2700 Hz. However, consonants are similar between whispered and neutral speech given the same speaker with the exception of liquid, glides and nasals. Considering the fact that voiced and unvoiced fricatives, affricatives and stops constitute the majority of consonants, our speaker ID system, which is introduced in the next section, only considers a specific compensation for the whisper vowels.

6. Speaker ID system

6.1. Methodology

The specific speaker ID task for whispered speech in this study assumes the absence of whispered adaptation data from *target speakers* and the availability of a small amount of whispered adaptation data from *non-target speakers*. However, results from Section 5 suggest that the difference between whispered and neutral speech is generally speaker dependent for vowels. This suggests that the direct estimation of a transformation given the *target speakers* whose whispered speech is not accessible is very challenging. This study avoids this test-phase estimation. Instead, we propose an estimation method to generate the pseudo-whisper features from neutral training features.

Given a small amount of whispered speech from *non-target speakers*, a speaker-independent whisper UBM can be trained. The goal here is to generate a pseudo-whisper feature corresponding to each given neutral feature using this whisper UBM. In this way, equal amounts of “whispered” and neutral data can be used to train a SMI-UBM. In Section 5.2, the statistical analysis shows that the difference between whispered and neutral vowels is generally speaker dependent. Therefore, given a neutral utterance from one speaker containing only vowels, we can use a global transformation to map the neutral features to whisper features for this speaker. The observations in Section 5.2 also suggest that the difference between whispered and neutral speech is similar across different phoneme contexts for vowels, therefore, the whisper UBM employed in this session is a GMM instead of HMM in Section 5.

This section considers three transformation formulations that model the difference between the neutral features and the whisper features: ConvTran, MLLR, and factor analysis (FA). In the ConvTran model, the whisper feature is assumed to be obtained by passing the neutral feature through a linear filter with additive noise in time domain, therefore the relation between whisperer and neutral features is non-linear in the MFCC domain as shown in Eq.

(2). The MLLR model, however, uses a linear transformation estimated at the maximum likelihood point to simulate the difference between whispered and neutral MFCC features. MLLR has been employed in many past studies for voice conversion (Dempster et al., 1977; Ye and Young, 2006) and adaptive training for speech recognition (Deguchi et al., 2010). It is also used in this study as a comparison to the ConvTran and FA model. In particular, we use the same affine transformation parameters to capture both the mean and the covariance differences between whisper and neutral speech as in constrained MLLR (CMLLR) (Deguchi et al., 2010). In the FA model, the principle components of the differences between whispered and neutral speech are extracted by projecting this difference on a pre-trained low-dimensional total variability space (Kenny et al., 2007). Compared to the CMLLR model, the ConvTran and FA models are have fewer parameters. However, the FA model requires training the low-dimensional total variability space that captures the variance of the difference between whisper and neutral features, while the ConvTran parameters are estimated for each utterance.

The above three modeling methods are essentially similar transformation estimation problems with different transformation formulations, where the transformation parameters are estimated iteratively with the expectation–maximization (EM) algorithm. The left side of Fig. 7 shows the general steps to obtain the final estimated transformation. After the transformation is obtained, the compensation will be conducted on the original neutral feature to obtain the pseudo-whisper feature. Eventually, both the neutral features and the pseudo-whisper features will be used to train a speech mode independent UBM as shown on the right side of Fig. 7. Details on the system framework will be included in Section 6.2.

6.1.1. ConvTran model

The ConvTran model employed in this section is essentially the same as the one employed in Section 5.2, but instead of estimating h given whisper features and a neutral HMM, here we will estimate h using neutral features and a whisper UBM. Therefore, the parameter update formulas based on the EM algorithm as described in Section 4 will be also employed here. Specifically, the whisper feature is assumed to be obtained by passing the neutral features through a linear filter with additive noise. This assumption is valid because only the smoothed spectral envelope is considered here. In the MFCC domain, this relationship can be presented as follows:

$$y_{\mathbf{ne}} = x_{\mathbf{wh}} + h + g(x_{\mathbf{wh}}, h, n), \tag{14}$$

$$g(x_{\mathbf{wh}}, h, n) = C\log(1 + \exp(C^{-1}(n - x_{\mathbf{wh}} - h))). \tag{15}$$

The above model is chosen for two reasons. First, due to the introduction of nonlinearity, the complexity of the parameters to be estimated decreases significantly, thus reducing the chance of overfitting given the limited amounts of adaptation data, as well as increasing the speed of adaptation. Second, considering that the differences between whispered and neutral vowels in the spectral envelope are mostly caused by formant and slope shifting, this model provides a reasonable way to capture aspects of the smoothed spectral envelope differences between whispered and neutral speech in the MFCC domain. After h is estimated, under the assumption of zero mean additive noise, a pseudo-whisper feature given the neutral feature can be obtained through:

$$x_{\hat{\mathbf{wh}},t} \approx y_{\mathbf{ne},t} - h. \tag{16}$$

In order to estimate h given the feature stream of a neutral utterance and a whisper UBM, the general VTS adaptation algorithm described in Section 4 is implemented as in Fig. 8.

In addition to the estimation model in Eq. (14), the implementation in this section differs from that in Section 5.1 as follows: only GMMs are considered here, which makes the calculation of $\gamma_{t,m}$ faster. Also, due to the duration difference between whispered and neutral speech, it is observed

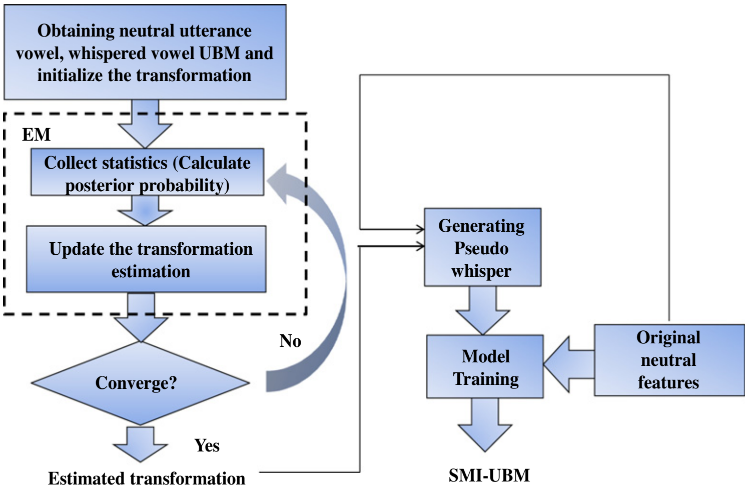
Obtaining neutral utterance vowel, whispered vowel UBM and initialize the transformation
EM
Collect statistics (Calculate posterior probability)
Update the transformation estimation
Converge?
No
Yes
Estimated transformation
Generating Pseudo whisper
Model Training
Original neutral features
SMI-UBM

Fig. 7. Transformation estimation and SMI-UBM model training.

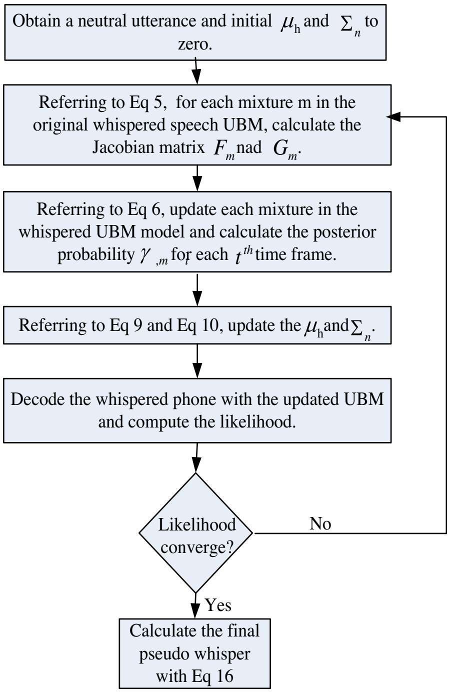
Obtain a neutral utterance and initial μ_h and Σ_n to zero.
Referring to Eq 5, for each mixture m in the original whispered speech UBM, calculate the Jacobian matrix F_m nad G_m.
Referring to Eq 6, update each mixture in the whispered UBM model and calculate the posterior probability $\gamma_{t,m}$ for each t^{th} time frame.
Referring to Eq 9 and Eq 10, update the μ_h and Σ_n.
Decode the whispered phone with the updated UBM and compute the likelihood.
Likelihood converge?
No
Yes
Calculate the final pseudo whisper with Eq 16

Fig. 8. Implementation of VTS adaptation algorithm for generating pseudo-whisper features.

that appending the delta coefficients will degrade performance, and thus only static features are considered here.

6.1.2. CMLLR model

Maximum likelihood linear transformation has been employed in Ye and Young (2006) for the purpose of voice conversion and in Deguchi et al. (2010) for compensation of acoustic differences introduced by various recording conditions in body-conducted speech conversation. As a common algorithm to model the mismatch introduced by speaker, background noise, or channel differences through a linear regression, it is also employed in this study as a comparison to the method described in Section 6.1.1. In the CMLLR context, the differences between whisper MFCC feature x_{wh} and neutral features y_{ne} are modeled as an affine transformation (A, b) as follow:

$$y_{ne} = Ax_{wh} - b. \tag{17}$$

Given Eq. (17), the relation between the mean and covariance of x_{wh} and y_{ne} can be represented as:

$$\mu_{\mathbf{ne}} = A\mu_{\mathbf{wh}} - b, \tag{18}$$

$$\Sigma_{\mathbf{ne}} = A\Sigma_{\mathbf{wh}}A^T. \tag{19}$$

By using the EM algorithm to iteratively maximize an auxiliary function, an estimation of A and b at the maximum likelihood point can be obtained (Gales, 1998). The pseudo-whisper feature can thus be obtained through:

$$\hat{x_{\mathbf{wh}}} = A^{-1}y_{\mathbf{ne}} + A^{-1}b. \tag{20}$$

Assuming that the dimension of the feature vector is $\mathcal{M}$, there will be a total of $\mathcal{M}$ ($\mathcal{M}$+ 1) parameters to be estimated. Therefore, to avoid the problem of overfitting, unlike estimating a transformation at utterance level as in the ConvTran model, a CMLLR is performed to obtain an estimation of A and b at speaker level, where all neutral utterances from a given speaker will be used to obtain a single global transformation. Compared with the ConvTran method in Section 6.1.1, the CMLLR method has the advantage that multiple transformations can be easily incorporated using methods such as regression class trees in order to capture the variability of the mismatch. However, because the acoustic analysis in Section 5 showed that the difference between whispered and neutral speech is generally speaker dependent, one global transformation is used for each speaker.

6.1.3. Factor analysis model

State-of-the-art performance in speaker ID is typically achieved with factor analysis (Kenny et al., 2007; Dehak et al., 2009; Lei and Hansen, 2009). In these systems, the speaker/channel variability is modeled through a low dimensional subspace. Most factor analysis based frameworks for speaker ID, including joint factor analysis (JFA) (Kenny et al., 2007) and total variability modeling (TVM) (Dehak et al., 2009), are based on probabilistic principle component analysis (PPCA) (Tipping and Bishop, 1999), where the principal eigenvectors are estimated with a finite set of speakers/channel training data. This approach supports estimation of eigenvectors from a relatively smaller training set (Kenny et al., 2005). In Section 6.1.1 and Section 6.1.2, the difference between whispered and neutral features was modeled with a convolutional filter and an affine transformation. In this section, the difference between whispered and neutral speech is projected onto a low dimensional subspace using factor analysis. The projection result is a latent vector used in an estimate of pseudo-whispered data.

In the context of FA, we use a similar modeling strategy as the TVM (Dehak et al., 2009). Instead of using a rectangular matrix T of low rank to capture the variability of speakers and channels, this study uses T to model the principle subspace for speakers and speech mode changes. In particular, the whisper UBM is represented by a supervector, which is simply the concatenation of the mean vectors given all mixtures of the UBM. A latent vector ω represents the projection of a given neutral utterance on the subspace T. Since the transformation is estimated in the supervector domain, a neutral utterance is represented by a supervector $M_{\mathbf{ne}}$ as:

$$M_{\mathbf{ne}} = m_{\mathbf{wh}} + T\omega, \tag{21}$$

where $m_{\mathbf{wh}}$ is the speaker-independent whisper supervector, T is the total variability space and ω is a random vector having a standard normal distribution $N(0, I)$. This model can be seen as a projection of the difference between speaker-independent whispered and speaker-dependent neutral speech into a low-dimensional total variability space. Therefore, the distribution of $M_{\mathbf{ne}}$ given ω can be seen as $N(m_{\mathbf{wh}}, TT^t)$.

Unlike the ConvTran model, where the value of h needs to be estimated, or the CMLLR model, where a linear transformation matrix needs to be estimated, in FA model, we will obtain the subspace T first on a development set. In this study, we employ 1 h neutral data from 20 female speakers in UT-VocalEffort II corpus without overlap with the NW28 set as the development set for T. Using this development set together with the same whisper UBM used in ConvTran and CMLLR, we use the EM algorithm to obtain the subspace T in the same way as in Dehak et al. (2009).

Next, given each neutral utterance in the training set and the obtained T, we will estimate the projection ω and eventually obtain the pseudo-whisper features. The estimation of ω also follows Dehak et al. (2009) by using the EM algorithm. After we obtain the ω, the pseudo-whisper vector is obtained by:

$$x_{\hat{\mathbf{wh}},t} = \sum_{k=0}^{M-1} \gamma_{t,k}(M_{\mathbf{ne}} - T\omega)_k, \tag{22}$$

where $\gamma_{t,k}$ is the posterior probability for the kth whispered UBM Gaussian mixture and $(M_{\mathbf{ne}} - T\omega)_k$ represents the kth mixture components in the supervector. In this study, the dimension of the latent factor is 5.

6.2. System

6.2.1. Baseline

As described in Section 3, the WH10 data set is used to train the whispered UBM in order to generate pseudo-whispered data. The NW28 set is used to train and test the speaker ID system. The feature parameters used in this study are 19-dimensional static Mel-frequency cepstral coefficients (MFCCs). All silence parts for whispered and neutral speech systems are first removed using a dynamic energy threshold that depends on the SNR of each particular sentence block sequence. The analysis frame length is 25 ms, with a 10 ms frame shift. For the baseline system, models for each speaker are obtained via MAP adaptation of a 64 mixture neutral UBM trained with an average of 4.5 min of neutral data per speaker from the set NW28. Fig. 9 shows the procedures for training/testing the baseline model. The whispered UBM for ConvTran, CMLLR and FA transformation trained using whisper data in the set WH10 has a fixed mixture size of 16.

6.2.2. ConvTran/CMLLR/FA based system

The model training procedures when ConvTran/CMLLR/FA models are incorporated are shown in Fig. 10. Given every neutral utterance, we will extract the vowel part, which will be used together with the whisper UBM to generate the pseudo-whisper features as described in Section 6.1. After we obtain the corresponding pseudo-whisper features for all given neutral features, we will further train a SMI-UBM. With this system, we can either use only neutral data to adapt this SMI-UBM as we do in the baseline system, or we can select some data from the pseudo-whisper feature pool and use them to adapt the SMI-UBM together with neutral data. The feature selection process is discussed in detail later in this section.

The vowel/consonant detection used to extract neutral vowels is implemented using two GMMs trained with neutral vowels and consonants respectively, where the vowels and consonants are obtained by using forced-alignment on TIMIT database. Given neutral speech data, each frame is tested against these two GMMs and tagged as the class that achieves the higher likelihood. Hence, for each neutral utterance, a neutral vowels set can be obtained and the ConvTran/CMLLR/FA based feature compensation will be only conducted on these neutral vowels.

Fig. 11a shows the log Mel power spectra versus time of a neutral utterance with all consonants and silence removed, since the compensation is only conducted on the vowel part of neutral features in our study. Fig. 11 shows the resulting pseudo-whisper log Mel power spectra via (b) ConvTran, (c) CMLLR and (d) FA model respectively. The log Mel power spectra of the real whispered vowels sequence (with consonants and silence removed) from the same speaker is shown in Fig. 12 for comparison. They are the same vowels in a different phoneme context.

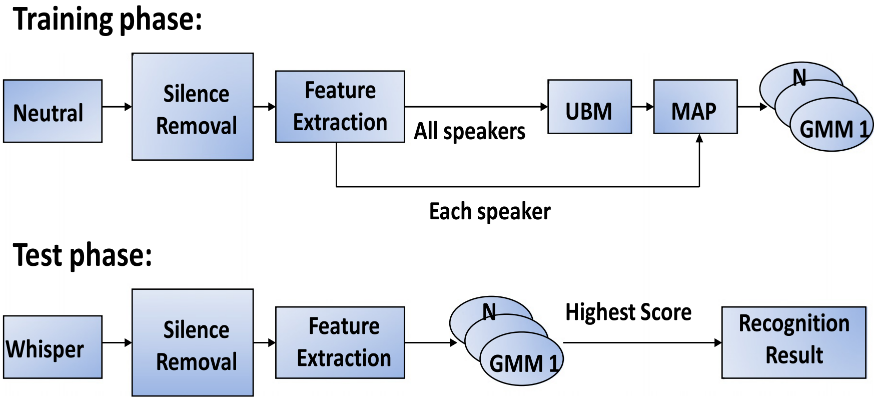

Fig. 9. Framework for baseline training/testing.

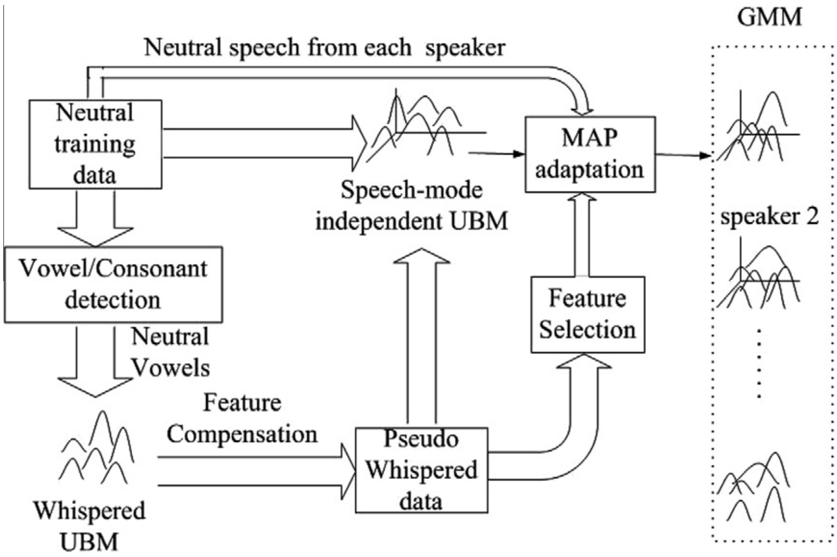

Fig. 10. System flow diagram for training GMMs.

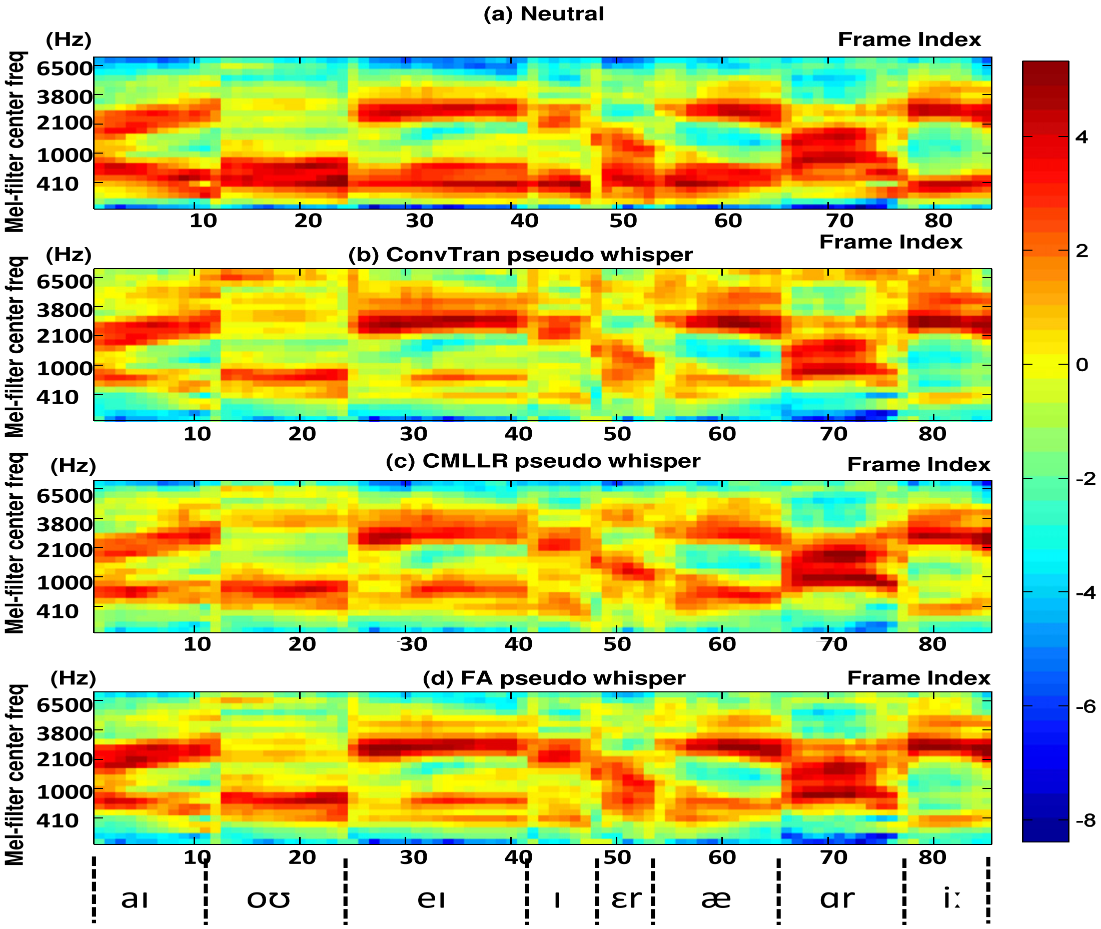

Fig. 11. (a) Is the log Mel power spectra of a neutral utterance with consonants (fricatives, affricatives, stops) removed, (b) is the corresponding pseudo-whispered log Mel power spectra obtained using ConvTran model, (c) is the corresponding pseudo-whispered log Mel power spectra obtained using CMLLR model and (d) is the corresponding pseudo-whispered log Mel power spectra obtained using FA model. (Note: 100 frames correspond to features for 1 second of speech duration with silence and consonant frames removed.)

Fig. 12 shows that, compared to CMLLR and FA, the pseudo-whisper features obtained from ConvTran model provide greater similarity to real whisper. For example, the energy below 1000 Hz is better suppressed.

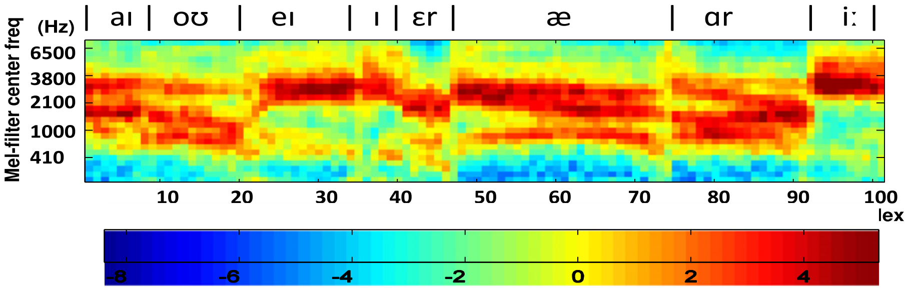

Fig. 12. Log Mel power spectra of some whispered vowels from the same speaker as in Fig. 11a.

Table 3
KL divergence between UBMs trained with different vowels.

KLD	real NE	real WH	ConvTran WH	CMLLR WH	FA WH
real WH	12.1	0	0.4	1.6	2.4
real NE	0	12.1	8.6	4.6	6.5

Components above 4000 Hz are properly emphasized. The bandwidth of formants in higher frequencies are increased. These differences were also observed in Ito et al. (2005) when comparing the acoustic properties of real whispered and neutral speech.

In order to quantitatively and statistically measure the distances among pseudo-whisper features, original neutral features, and real whisper features, the Kullback Leibler (KL) divergence (Kullback et al., 1968) is applied to compare the distance of UBMs trained with different sets of features. The asymmetric property of the KL divergence is resolved by taking the mean of $D(P\|Q)$ and $D(Q\|P)$. A smaller value of KL divergence indicates two similar probability distribution. UBMs considered here include: one UBM trained with real whispered data from the NW28 set; one UBM trained with real neutral data from the NW28 set, and three UBMs trained with pseudo-whispered data generated from neutral data in the NW28 set through ConvTran, CMLLR, and FA model, respectively. Table 3 lists results. Compared to CMLLR and FA, the UBMs obtained through ConvTran pseudo-whisper features shows a closer distance (0.4) to UBMs trained with real whisper features and a farther distance (8.6) to UBMs trained with real neutral features. The results quantitatively confirm the observation in Fig. 11 and show that the pseudo-whisper in fact moves toward whispered speech.

After pseudo-whispered vowels are generated through ConvTran/CMLLR/FA, equal amounts of neutral and pseudo-whispered vowels are available for model training. In order to balance the distribution of phonemes, the neutral consonants from the vowel/consonant detection are also used to train the SMI-UBM model. The complexity of the UBM is also doubled to 128 mixtures to account for the increased diversity of training data.

The purpose of the feature selection procedure in Fig. 10 is to select pseudo-whisper features that are similar to the real whispered data for subsequent MAP adaptation. The candidates for feature selection are obtained from the pseudo-whispered vowels generated from the same speaker. For example, in order to obtain the GMM for Speaker 1, feature selection only considers the pseudo-whispered vowels obtained from Speaker 1’s neutral data. The criterion of selection is the correctness of recognition by the neutral trained GMMs. For example, given two pseudo-whispered vowels from Speaker 1: **wh**A and **wh**B, they will be tested against the GMMs obtained from MAP adaptation using only neutral data. If **wh**A is recognized as Speaker 2 and **wh**B is recognized as Speaker 1, **wh**B will be selected for MAP adaptation to obtain Speaker 1’s GMM. If none of the pseudo-whispered vowels are correctly recognized, those that achieve the highest rank will be chosen. The available amount of pseudo-whispered adaptation data is also under the constraint that the average must be equal among all the speakers. This prevents the amount of adaptation data for any given speaker from being much greater than any other. This study uses on average 5–10 s pseudo-whispered data per each speaker in the adaptation step.

A minimum mean square error (MMSE) criterion between the pseudo-whispered vowels and the conditional expectation of it given the neutral model was also considered. However, this criterion resulted in poorer overall performance, and hence will not be discussed here. For simplicity, feature selection will be referred to as FS for the remainder of this study.

6.2.3. Experimental results

Given the proposed ConvTran/CMLLR/FA based training procedure, the testing phase employs the same procedure as that used for the baseline system. A total of 961 whispered utterances from the NW28 set are employed for recognition. Table 4 summarizes the data used for UBM training and MAP adaptation in all systems. Cepstral mean normalization (CMN) is a simple method for suppressing microphone and channel effects, so it is also employed for comparison. However, the resulting accuracy is only 23.93%.

Fig. 13 shows that the baseline system provides an accuracy of 79.29% (i.e., speaker ID models trained with neutral speech; all test data are whispered speech). When we use

Table 4
Data for training UBM and speaker dependent GMM. NE represents neutral speech; WH represents whispered speech; FS represents feature selection. The corpus name in the bracket after NE or WH indicates where are the whispered and neutral data from.

System/data	Training data for UBM	Adaptation data for MAP
Baseline	NE(NW28)	NE(NW28)
Whispered UBM	NE(NW28) + WH(WH10)	NE(NW28)
CMLLR wo FS	NE(NW28) + pseudo WH	NE(NW28)
FA wo FS	NE(NW28) + pseudo WH	NE(NW28)
ConvTran wo FS	NE(NW28) + pseudo WH	NE(NW28)
CMLLR with FS	NE(NW28) + pseudo WH	NE(NW28) + pseudo WH (5–10 s/spkr)
FA with FS	NE(NW28) + pseudo WH	NE(NW28) + pseudo WH (5–10 s/spkr)
ConvTran wth FS	NE(NW28) + pseudo WH	NE(NW28) + pseudo WH (5–10 s/spkr)

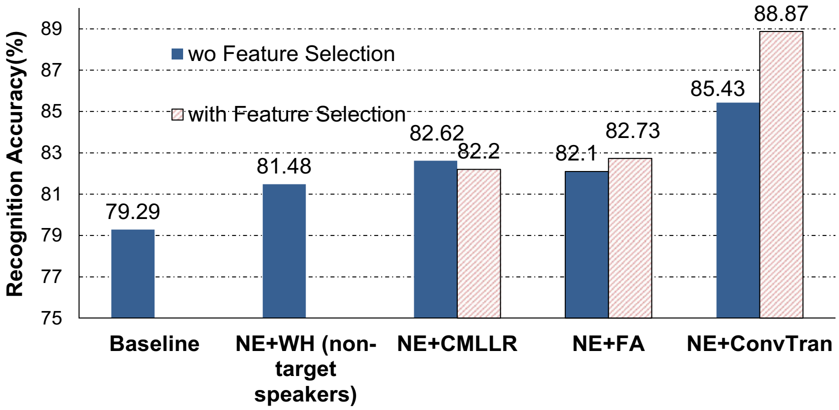

Fig. 13. Recognition results for closed set speaker ID using whispered test data, where speaker ID models are either trained with neutral, neutral + whisper, or neutral + various pseudo-whisper schemes.

half of the neutral data from the NW28 set to train the speaker ID model, and another half of the neutral data from the NW28 set to test the obtained model, an accuracy of 99.06% is achieved. The results confirm the degradation in speaker ID performance caused by the whisper/neutral mismatched training/testing condition. When we combine all the whispered data from the WH10 set with all the neutral speech from NW28 to train a UBM using only neutral data for MAP adaptation, a performance of 81.48% is obtained. This result suggests that if some whispered data is incorporated in training the UBM, even if the whispered speech is not from the "target" speakers, some level of whispered speech style is captured to improve the performance.

When all pseudo-whispered data obtained from the (i) ConvTran, (ii) CMLLR or (iii) FA models is employed to train the SMI-UBM along with the available neutral speech, a closed-set speaker ID performance of 85.43%, 82.10% and 82.62% is obtained respectively. As described in Section 6.1, in the ConvTran model, the estimation is obtained at the utterance level where each neutral vowel stream contains about 3–10 s data. In CMLLR model, we estimate an affine transformation per speaker using all neutral data from the speaker. Thus in the NW28 set, there are a total of 28 transformations needed to be obtained, where each speaker has an average of 3 min neutral vowel data. In the FA model, the transformation is represented as ω in Eq. (22) and the estimation is obtained at the utterance level as the same as the ConvTran model.

Although, the model complexity for the FA model at the estimation step is also small compared with the CMLLR model, the FA model requires a fair amount of data in order to obtain a valid low dimensional subspace at the training steps. The ConvTran model does not require any data for pre-training.

When combined with feature selection, the highest performance of 88.87% is achieved with the ConvTran model using 5–10 s of pseudo-whispered data selected for each speaker as described in Section 6.2. This represents a significant relative improvement of +46.6% in closed-set speaker recognition accuracy and demonstrates the effectiveness of the feature selection strategy. The system based on ConvTran outperforms both the FA and CMLLR systems, which indicates that the generated pseudo-whispered features from the ConvTran model keep more speaker-dependent information than those generated from FA and CMLLR.

We used the model from ConvTran pseudo-whisper with 88.87% accuracy to test 953 neutral utterances not seen from the training phase. An accuracy of 99.37% is achieved in this case, which demonstrates that the proposed technique does not reduce system performance on neutral speech.

7. Conclusion

The goal of this study has been to develop a robust speaker ID system which can provide sustained performance for whispered speech in the absence of any speaker-dependent whispered adaptation data. An acoustic analysis was conducted first in order to develop an efficient model training/adaptation method and those analysis results suggested that the difference between whispered and neutral speech is generally consistent across speakers, especially beyond 4 kHz. Shifts in spectral tilt due to whisper of consonants were shown to differ across five consonant categories. It was also observed that the differences between whispered and neutral speech focus on liquids, glides, and nasals.

Based on results from the acoustic analysis, a new system framework was proposed that resulted in a speech mode independent model without the requirement of a parallel data collection for whispered and neutral speech. Three transformation models were employed in this study, including ConvTran, CMLLR, and FA, where the ConvTran model provided an overall better quality of pseudo-whispered features based on KL divergence measurement. With the highest closed-set speaker ID accuracy of 88.87%, a relative improvement of 46.6% was achieved. The proposed system also retains the neutral test data performance with an 99.37% accuracy, and retains the conventional test procedure for speaker recognition systems. Thus no additional data processing or calculation is required during the test phase.

A similar method can be helpful for speech transformation from whispered speech to neutral speech as well, which is usually implemented by using a codebook. The study has therefore offered a viable approach for closed-set speaker ID of whispered speech when whispered training data is not available for the target speaker set, and a possible method to balance model training when insufficient data is available under specific speaking conditions.

Acknowledgement

The authors would like to thank our colleagues at CRSS-UTD Keith Godin and Yun Lei, for their useful discussions and software development support for this study.

References

Bou-Ghazale, S., Hansen, J.H.L., 1998. Stress perturbation of neutral speech for synthesis based on hidden Markov models. IEEE Transactions on Speech and Audio Processing 6, 201–216.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representation for monosyllabic word recognition in continuous spoken sentences. IEEE Transactions on Acoustics Speech and Signal Processing 28 (4), 357–366.

Deguchi, D., Doi, H., Toda, T., Saruwatari, H., Shikano, K., 2010. Acoustic compensation method for accepting different recording devices in body-conducted voice conversion. In: APSIPA ASC, Biopolis, Singapore, pp. 502–505.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., Dumouchel, P., 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Interspeech, Brighton, UK, pp. 1559–1562.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39 (1), 1–38.

Deng, L., Droppo, J., Acero, A., 2004. Estimating spectrum of speech under the presence of noise using a joint prior of static and dynamic features. IEEE Transactions on Speech and Audio Processing 12 (3), 218–233.

Eklund, I., Traunmuller, H., 1996. Comparative study of male and female whispered and phonaed versions of the long vowels of Swedish. Phonetica 54, 1–21.

Fan, X., Hansen, J.H.L., 2008. Speaker identification for whispered speech based on frequency warping and score competition. In: INTERSPEECH, Brisbane, Australia, pp. 1313–1316.

Fan, X., Hansen, J.H.L., 2009. Speaker identification for whispered speech using modified temporal patterns and MFCCs. In: INTERSPEECH, Brighton, UK, pp. 896–899.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language 12, 75–98.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V., 1993. TIMIT acoustic-phonetic continuous speech corpus. In: Linguistic Data Consortium, Philadelphia, USA, 1993.

Gavidia-Ceballos, L., Hansen, J.H.L., 1996. Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. IEEE Transactions on Biomedical Engineering 43, 373–383.

Ito, T., Takeda, K., Itakura, F., 2005. Analysis and recognition of whispered speech. Speech Communications 45, 139–152.

Jin, Q., Jou, S.S., Schultz, T., 2007. Whispering speaker identification. In: IEEE International Conference on Multimedia and Expo, Beijing, China, pp. 1027–1030.

Jovicic, S.T., 1998. Formant feature differences between whispered and voiced sustained vowels. Acustica-Acta 84, 739–743.

Jovicic, S.T., 1998. Formant feature differences between whispered and voiced sustained vowels. Acustica-Acta 84 (4), 739–743.

Kallail, K.J., Emanuel, F.W., 1984. Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. Journal of Speech Hearing Research 27, 245–251.

Kenny, P., Boulianne, G., Dumouchel, P., 2005. Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing 13, 345–354.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Speech and Audio Processing 15, 1435–1447.

Kullback, S., 1968. Information Theory and Statistics. Dover Publications, Inc., Mineola, NY.

Lei, Y., Hansen, J.H.L., 2009. Factor analysis-based information integration for arabic dialect identification. In: ICASSP, Taipei, China, pp. 4337–4340.

Li, J., Yu, D., Deng, L., Gong, Y., Acero, A., 2009. A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions. Computer Speech and Language 23, 389–405.

Matsuda, M., Kasuya, H., 1999. Acoustic nature of the whisper. In: EUROSPEECH, pp. 133–136.

Meyer-Eppler, W., 1957. Realisation of prosodic features in whispred speech. Journal of the Acoustical Society of America 29, 104–106.

Moreno, P.J., Raj, B., Stern, R.M., 1996. A vector Taylor series approach for environment-independent speech recognition. In: ICASSP, pp. 733–736.

Morris, R.W., Clements, M.A., 2002. Reconstruction of speech from whispers. Medical Engineering and Physics 24, 515–520.

Thomas, I., 1969. Perceived pitch of whispered vowels. Journal of the Acoustical Society of America 46, 468–470.

Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. Journal of the Royal Statistical Society, 611–622.

Ye, H., Young, S., 2006. Quality-Enhanced Voice Morphing Using Maximum Likelihood Transformations, 1301–1312.

C. Zhang, J.H.L. Hansen, Analysis and classification of speech mode: whisper through shouted. In: INTERSPEECH 2007, August 2007, Antwerp, Belgium, pp. 2289–2292.

Zhang, C., Hansen, J.H.L., 2009. Advancement in whisper-island detection with normally phonated audio streams. In: INTERSPEECH, Brighton, UK, pp. 860–863.