Speaker Identification for Whispered Speech Using A Training Feature Transformation From Neutral To Whisper

Xing Fan and John H.L. Hansen

Center for Robust Speech Systems (CRSS)
University of Texas at Dallas, Richardson, Texas 75083, USA
xxf064000@utdallas.edu, john.hansen@utdallas.edu

Abstract

A number of research studies in speaker recognition have recently focused on robustness due to microphone and channel mismatch(e.g., NIST SRE). However, changes in vocal effort, especially whispered speech, present significant challenges in maintaining system performance. Due to the mismatch spectral structure resulting from the different production mechanisms, performance of speaker identification systems trained with neutral speech degrades significantly when tested with whispered speech. This study considers a feature transformation method in the training phase that leads to a more robust speaker model for speaker ID with whispered speech. In the proposed system, a Speech Mode Independent (SMI) Universal Background Model (UBM) is built using collected real neutral features and pseudo whispered features generated with Vector Taylor Series (VTS), or via Constrained Maximum Likelihood Linear Regression (CMLLR) model adaptation. Text-independent closed set speaker ID results using the UT-VocalEffort II corpus show an accuracy of 88.87% using the proposed method, which represents a relative improvement of 46.26% compared with the 79.29% accuracy of the baseline system. This result confirms a viable approach to improving speaker ID performance for neutral and whispered speech mismatched conditions.
Index Terms: whispered speech, speech identification

1. Introduction

Whispered speech is an alternative vocal effort style from neutral speech which is employed between speakers when conveying personal information. For example, when making a hotel/car reservation over a cell phone in a public area, a speaker may whisper in order to provide information related to credit card info, billing address, or phone number. Individuals, such as aphonic patients, heavy smokers, also employ whispered speech as their primary oral communication method. Compared with neutral speech, whispered speech has no fundamental frequency due to the absence of voiced excitation. The differences in production is also reflected in formant shifting and slope changes [1,2,3]. Therefore, the performance of speaker ID systems trained with neutral speech degrades significantly when tested with whispered speech.

Past work on speaker ID for whispered speech can be grouped under two main categories: front-end processing[5,6] and model adaptation[7]. Improvements in performance have resulted with both categories. However, a new front-end processing method will involve feature re-extraction and model re-training for neutral speech, which introduces more computational requirement as well as a potential decrease in performance for neutral speech. For model adaptation, a simple Maximum a posteriori (MAP) adaptation system can provide satisfactory performance under the prerequisite of a fair amount of speaker-dependent (SD) whispered adaptation data. However, whispered adaptation data from test speakers is generally not available in real applications. Even though it is possible to collect extra whispered data from other speakers, due to the fact that the amount of real whispered data is usually much smaller compared with available neutral data, it is still very difficult to build a reasonable starting SMI-UBM.

This study first proposes a method based on VTS/CMLLR to generate pseudo speaker-dependent whispered features given neutral features. Those features are further employed to train a SMI-UBM, which will include equal amounts of information form both neutral and whisper. Also, because the proposed method keeps some level of speaker-dependent information in the resulting pseudo whispered features, after a SMI-UBM is trained, a SD model can be further obtained by adaptation of the UBM with both neutral and selected pseudo whisper features. As such, there are two assumptions for applying the proposed method. First, the differences between whispered and neutral speech in vowels are similar given the same speaker[8]. Therefore, it is reasonable to estimate an overall transformation parameters set for each utterance. Second, an extra small amount of whispered data set is available, whose speakers are not required to be included in the whispered test set. This assumption is easy to satisfy in a real application. The remainder of this paper is organized as follows: Sec.2 introduces the method of VTS and CMLLR based feature compensation for the training phase. Sec.3 presents an introduction to the corpus and provides a description of the proposed overall system and experimental results. Conclusions are drawn in Sec. 4.

2. Training Data Preparation

Given a small amount of whispered speech, a whispered UBM can be trained. By using the obtained speaker-independent (SI) whispered UBM combined with model adaptation, the goal here is to generate a pseudo whispered feature corresponding to each given neutral feature. A similar perturbation concept was previous formulated for spectral structure, fundamental frequency, and duration properties for stressed emotional speech recognition [13]. In this study, equal amounts of whispered and neutral data can be obtained and a SMI-UBM can be balanced trained afterwards. Two adaptation methods are considered here respectively: VTS and CMLLR.

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.

2.1. VTS based Adaptation

In this section, the neutral feature is assumed to be obtained by passing the whispered feature through a linear filter with additive noise. This assumption is valid since only the smoothed spectral envelope is considered here. In the MFCC domain, this relation can be presented as follow:

$$\mathbf{ne} = \mathbf{wh} + h + g(\mathbf{wh}, h, n), \tag{1}$$

$$g(\mathbf{wh}, h, n) = C\log(1 + \exp(C^{-1}(n - \mathbf{wh} - h))). \tag{2}$$

where C^{-1} is the pseudo inverse of the DCT matrix. This model is chosen for two main 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 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 of whispered speech in the MFCC domain. The noise distortion n in Eq. (1) is assumed to have a Gaussian distribution with zero mean μ_n and a diagonal covariance matrix Σ_n. The filter h is assumed to be a fixed vector for a given neutral utterance. To resolve the introduction of nonlinearity, VTS based model adaptation, which has previously been considered for robust speech recognition system[9,10], is applied here with modifications.

2.1.1. Parameter Estimation

After applying a first order VTS approximation to Eq. (1) around the mean vector of the whispered UBM $\mu_{\mathbf{wh}}$, h, and μ_n, we have,

$$\begin{aligned}\mathbf{ne} &\approx \mu_{\mathbf{wh}} + h + g(\mu_{\mathbf{wh}}, h) \\ &+ G(wh - \mu_{\mathbf{wh}}) + G(h - \mu_{\mathrm{h}}) + F(n - \mu_n),\end{aligned} \tag{3}$$

where,

$$\begin{aligned}&\frac{\partial \mathbf{ne}}{\partial \mathbf{wh}}|_{\mu_{\mathbf{wh}},h} = \frac{\partial \mathbf{ne}}{\partial h}|_{\mu_{\mathbf{wh}},h} = G \\ &\frac{\partial \mathbf{ne}}{\partial n}|_{\mu_{\mathbf{wh}},h} = I - G = F \\ &G = C \cdot diag\{\frac{1}{1 + \exp(C^{-1}(-\mu_{\mathbf{wh}} - h))}\} \cdot C^{-1},\end{aligned} \tag{4}$$

where $diag$ stands for a diagonal matrix with its diagonal component value equal to the value of the vector in the argument. By taking the expectation and variance operation of both sides of Eq.(3), we have,

$$\begin{aligned}\mu_{\mathbf{ne}} &\approx \mu_{\mathbf{wh}} + h + g(\mu_{\mathbf{wh}}, h) \\ \Sigma_{\mathbf{ne}} &\approx G\Sigma_{\mathbf{wh}}G^t + F\Sigma_n F^t\end{aligned} \tag{5}$$

The Expectation and Maximization (EM) algorithm is employed iteratively to estimate h. Given a neutral utterance $\mathbf{ne}$, the auxiliary function is as follow:

$$Q(\lambda|\overline{\lambda}) = \sum_t \sum_m \gamma_{t,m} \log p(\mathbf{ne}_t|m, \lambda), \tag{6}$$

where $p(\mathbf{ne}_t|m, \lambda) \sim \mathcal{N}(\mathbf{ne}_t; \mu_m, \Sigma_m, \omega_m)$ and $\gamma_{t,m}$ is the posterior probability of the m^{th} Gaussian mixture in the updated whispered UBM for the t^{th} frame in $\mathbf{ne}$. In the M-step, Q is maximized by taking the derivative with respect to h. By setting the derivative to zero, the update formula for h is obtained as:

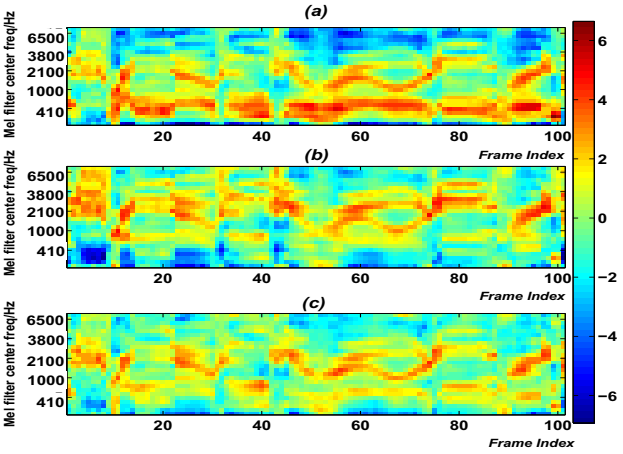

Figure 1: *Fig. 1(a) is the log Mel power spectra of a neutral utterance with consonants removed. Fig. 1(b) is the corresponding pseudo whispered log Mel power spectra obtained using VTS adaptation. Fig. 1(c) is the corresponding pseudo whispered log Mel power spectra obtained using CMLLR adaptation.*

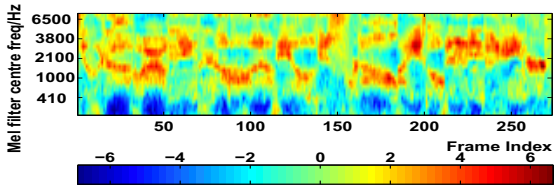

Figure 2: *Log Mel power spectra of some whispered vowels from the same speaker as in Fig. 1(a).*

$$\begin{aligned}h = h_0 + \{\sum_t \sum_m \gamma_{t,m} G_m^T \Sigma_{\mathbf{ne},m}^{-1} G_m\}^{-1} \\ \{\sum_t \sum_m \gamma_{t,m} G_m^T \Sigma_{\mathbf{ne},m}^{-1} [\mathbf{ne}_t - \mu_{\mathbf{ne},m} - h_0 - g(\mu_m, h_0)]\}\end{aligned} \tag{7}$$

The Σ_n is updated with Newton's method as:

$$\Sigma_n = \Sigma_{n,0} - (\frac{\partial^2 Q}{\partial^2 \Sigma_n})^{-1}(\frac{\partial Q}{\partial \Sigma_n}) \tag{8}$$

After h is estimated, under the assumption of zero mean additive noise, a pseudo whispered feature given the neutral feature can be obtained through:

$$\hat{\mathbf{wh}}_t \approx \mathbf{ne}_t - h \tag{9}$$

2.1.2. Algorithm Implementation

The procedures for obtaining a pseudo whispered feature given a neutral feature are summarized as follow:

1. Obtain a neutral utterance and initialize h and Σ_n to zero
2. For each mixture m in the original whispered speech UBM, calculate the Jacobian matrix G_m and F_m according to Eq. (4)

3. Update each mixture in the whispered UBM model according to Eq. (5) and calculate the posterior probability $\gamma_{t,m}$ for each t^{th} time frame.
4. Update h and Σ_n according to Eq. (7,8)
5. Decode the utterance with the updated UBM and calculate the likelihood. If the likelihood converges, record the estimated h. Otherwise, repeat the process by going back to Step 2.
6. Calculate the final pseudo whisper $\hat{\mathbf{wh}}_t$ with Eq. (9).

2.2. CMLLR based Adaptation

CMLLR allows estimated transformation applied in feature domain by constraining the variance transforms corresponding to the mean transformation [11]. Since, an affine transformation between the adaptation data and model parameters is estimated by maximizing the expected likelihood in CMLLR, it is also employed here as a comparison to the nonlinear-model described in Sec.2.1. In the context of CMLLR, the relation between whispered vowels $\mathbf{wh}$ and neutral vowels $\mathbf{ne}$ is modeled as:

$$\mu_{\mathbf{ne}} = A\mu_{\mathbf{wh}} - b; \tag{10}$$
$$\Sigma_{\mathbf{ne}} = A\Sigma_{\mathbf{wh}}A^T; \tag{11}$$

By using the EM algorithm to iteratively maximize the auxiliary function as in Eq. (12), an estimation of A and b in Eq. (10) can be obtained.

$$Q(\lambda|\bar{\lambda}) = \sum_t \sum_m \gamma_{t,m} \log p(\mathbf{ne}_t|A, b, m, \lambda) \tag{12}$$

The pseudo whispered feature can thus be estimated as:

$$\hat{\mathbf{wh}}_t = A^{-1}\mathbf{ne}_t + A^{-1}b \tag{13}$$

Assuming 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 independent assumption between utterances in the VTS based adaptation, a supervised CMLLR is performed to obtain an estimation of A and b for each speaker in the neutral set. Recognition experiment also confirms its advantage over an unsupervised CMLLR in Sec. 3.3. The unsupervised CMLLR is referred to simply as CMLLR for remainder of this study.

Fig.1(a) shows the log Mel power spectra of a neutral utterance with consonants and silence removed. Fig.1(b) and (c) shows the resulting pseudo whisper via VTS and supervised CMLLR adaptation respectively. The log Mel power spectra of some real whispered vowels from the same speaker is provided in Fig. 2 as well for comparison(different phoneme context). Fig. 2 shows that compared with supervised CMLLR, the pseudo whispered features obtained from VTS adaptation provide more similarities with real whisper. For example, the energy below 1000 Hz is well suppressed. Information beyond 4000 Hz is properly emphasized. The bandwidth of formants in the higher frequency is expended and some formants in lower frequency (100-1500 Hz) are also shifted to higher frequency. All of those differences are also observed in [1] when comparing the acoustic properties between real whispered and neutral speech.

3. System and Experimental Results

3.1. Corpus

The UT-VocalEffort II corpus developed in [12] is employed in this study. Whispered and neutral speech from 28 native American English female subjects are chosen for a closed-set speaker

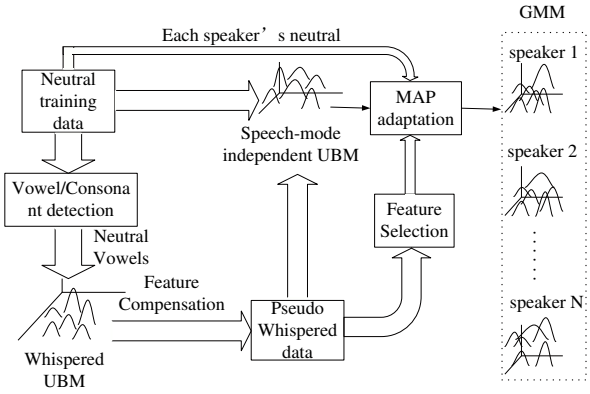

Figure 3: *System flow diagram for training GMMs.*

ID recognition task, where each speaker has an average neutral training data of 4.5 minutes and an average 34 whispered test utterances ranging from 1-3 second. Another 10 different female speakers' whispered speech, which is a total of 10 minutes whispered data, are employed for training the whispered UBM for VTS and CMLLR adaptation. For simplicity, we refer to the 28 speakers set as NW28 and the 10 speakers set as WH10 for the rest part of this paper. From [12], it is also noted 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 sections. Speech data was digitized using a sample frequency of 16 kHz, with 16 bits per sample.

3.2. System

3.2.1. Baseline 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. Due to the duration difference between whispered and neutral speech, it is observed that appending the delta coefficients will degrade the performance, thus they are not considered here. The analysis frame length is 25 ms, with a 10 ms frame shift. For the baseline system, models for each speaker is obtained by MAP adaptation of a 64 mixture neutral UBM trained with all the available neutral data from the 28 speakers. The whispered UBM for VTS and CMLLR adaptation has a mixture of 16.

3.2.2. VTS/CMLLR Based System

The model training procedures when VTS/CMLLR adaptation is incorporated are shown in Fig. 3. Since the difference between whispered and neutral speech mainly exists in vowels, and it was shown earlier that the consistency of this difference [8], a vowel/consonant detection is necessary to ensure a good source for further parameter estimation. The detection is implemented here by using vowel and consonant GMMs followed with likelihood comparison. Given neutral data, each frame will be tested against these two GMMs and tagged as the class that achieves higher likelihood. Hence, for each neutral utterance, a neutral vowels set can be obtained and the VTS/CMLLR based feature compensation is only conducted on those neutral vowels.

After the pseudo whispered vowels are obtained through

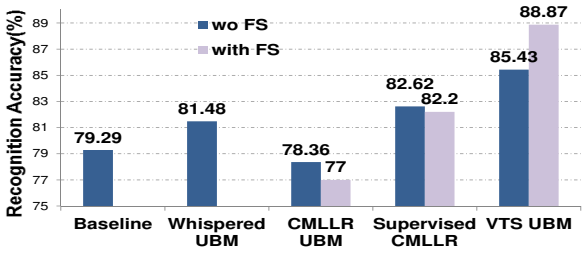

Figure 4: *Recognition results for closed set speaker ID using whispered test data.*

VTS/CMLLR as described in Sec.2, Equal amounts of neutral and pseudo whispered vowels are available. In order to balance the distribution of phonemes, the neutral consonants from the vowel/consonant detection are also fed to train the SMI-UBM model. The new UBM's mixture number is also doubled to 128 as a result of the increased training data.

The purpose of the feature selection procedure in Fig. 3 is to select pseudo whispered features that are similar to the real whispered data for the following MAP adaptation. The candidates of feature selection are from pseudo whispered vowels generated from the same speaker. For example, in order to obtain the GMM for Speaker 1, the feature selection only considers pseudo whispered vowels obtained from Speaker 1's neutral data. The criterion of selection here is simply 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 GMMs obtained from MAP only using 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 model. If none of the pseudo whispered vowels are correctly recognized, those achieve the highest rank will be chosen. The available amount of pseudo whispered adaptation data is also under the constraint that it is average equal among all speakers. Another criterion: the Minimum mean square errors between the pseudo whispered vowels and the conditional expectation of it given the neutral model was also considered, which resulted in poorer performance, hence will not be discussed here. For simplicity, feature selection will be referred to as FS for the rest of this study.

3.3. Experimental Results

Given the proposed VTS/CMLLR based training procedure, the testing phase keeps the same procedure as the baseline system. A total of 961 whispered utterances from the NW28 set are employed for recognition. Fig. 4 summarizes the results. Cepstral mean normalization (CMN) is a simple and efficient method for removing noise and channel effect, so it is also employed. However, the resulting accuracy is only 23.93% in our experimental set-up, which means the whispered/neutral speech mismatch cannot be simply removed with a mean normalization.

From Fig. 4, it can be observed that the baseline system provides an accuracy of 79.29%. 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 even the whispered data for training the UBM is not from the same speakers for testing; the incorporating of general whispered acoustic properties into the model helps for improving performance. When all the pseudo whispered data obtained from VTS or supervised-CMLLR adaptation are employed to train the SMI-UBM along with the available neutral speech, a performance of 85.43% and 82.62% is obtained respectively. Unsupervised CMLLR, however, causes a degradation to 78.36% when compared with the baseline because of the introduction of the overfitting problem due to short adaptation data. When combined with feature selection, the highest performance of 88.87% is achieved by VTS adaptation with 5-10 seconds pseudo whispered data selected for each speaker as described in Sec.3.2, which represents a significant relative improvement of +46.6% in accuracy.

4. Conclusions

In this paper, a new system framework was proposed that resulted in a mode independent model based on VTS/CMLLR adaptation without large labor of data collection for whispered speech. With the highest accuracy of 88.87%, a relative improvement of 46.6% was achieved. The proposed method also maintains the conventional test procedure for speaker recognition systems, thus no additional data transportation 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 hard codebook.

5. References

[1] T. Ito, K. Takeda and F. Itakura, "Analysis and Recognition of Whispered Speech," Speech Communication pp.139-152, 2005

[2] S. T. Jovicic, "Formant Feature Differences between Whispered and Voiced Sustained Vowels," Acustica-acta, pp.739-743, 1998

[3] M. Matsuda and H. Kasuya, "Acoustic Nature of the Whisper," EUROSPEECH, pp.137-140, 1999

[5] X. Fan and J. H. L. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition,"INTERSPEECH, pp.1313-1316, 2008

[6] X. Fan and J.H.L. Hansen, "Speaker Identification for whispered speech using modified temporal patterns and MFCCs," INTERSPEECH 2009

[7] Q. Jin, S. S. Jou and T. Schultz, "Whispering Speaker Identification," IEEE International Conference on Multimedia and Expo, 2007

[8] X. Fan K. Godin and J. H. L. Hansen, "Acoustic Analysis of whispered speech for phoneme&speaker dependentcy," sumbmitted to INTERSPEECH, 2011

[9] Moreno. P.J, Raj. B, Stern, R.M, "A vector Taylor series approach for environment-independent speech recognition" ICASSP, pp.733-736, 1996

[10] L. Deng, J. Droppo and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features", IEEE Trans on SAP, pp 218-233, 2004

[11] M.J.F. Gales and P.C. Woodland, "Mean and Variance Adaptation within the MLLR Framework",Computer Speech & Language, pp. 249-264, 1996

[12] C. Zhang and J.H.L. Hansen, "Advancement in whisper-island detection with normally phonated audio streams,"INTERSPEECH, pp.860-863, 2009.

[13] S. Bou-Ghazale, J.H.L. Hansen, "Stress Perturbation of Neutral Speech for Synthesis based on Hidden Markov Models," IEEE Trans on SAP, pp.201-216, 1998