Speaker Identification Within Whispered Speech Audio Streams

Xing Fan and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Whisper is an alternative speech production mode used by subjects in natural conversation to protect the privacy. Due to the profound differences between whisper and neutral speech in both excitation and vocal tract function, the performance of speaker identification systems trained with neutral speech degrades significantly. In this paper, a seamless neutral/whisper mismatched closed-set speaker recognition system is developed. First, performance characteristics of a neutral trained closed-set speaker ID system based on an Mel-frequency cepstral coefficient–Gaussian mixture model (MFCC-GMM) framework is considered. It is observed that for whisper speaker recognition, performance degradation is concentrated for only a subset of speakers. Next, it is shown that the performance loss for speaker identification in neutral/whisper mismatched conditions is focused on phonemes other than low-energy unvoiced consonants. In order to increase system performance for unvoiced consonants, an alternative feature extraction algorithm based on linear and exponential frequency scales is applied. The acoustic properties of misrecognized and correctly recognized whisper are analyzed in order to develop more effective processing schemes. A two-dimensional feature space is proposed in order to predict on which whispered utterances the system will perform poorly, with evaluations conducted to measure the quality of whispered speech. Finally, a system for seamless neutral/whisper speaker identification is proposed, resulting in an absolute improvement of 8.85%–10.30% for speaker recognition, with the best closed set speaker ID performance of 88.35% obtained for a total of 961 read whisper test utterances, and 83.84% using a total of 495 spontaneous whisper test utterances.**

***Index Terms*—Mel-frequency cepstral coefficient (MFCC), robust speaker verification, speaker identification, vocal effort, whispered speech.**

I. Introduction

WHISPERED speech is a natural mode of speech production that may be employed in public situations to protect the content of speech information. When speaking on a cell phone in a public setting, a speaker may prefer to whisper when providing their credit card number, bank account number, or other personal information. When customers make hotel, flight, or car reservations by telephone, they may whisper to provide information regarding their date of birth, credit card information, and billing address. Doctors may whisper if it is necessary to discuss patient medical records in public settings to maintain patient confidentiality. Aphonic individuals, as well as those with low vocal capability, such as heavy smokers, speak in a whisper mode as their primary form of oral communication.

The speech spectra will reflect significant differences between whisper and neutral speech[1] production. Differences include a complete loss of voiced excitation structure, and a shift in formant center frequencies in low frequency regions [1]–[5]. Zhang and Hansen [1] showed that changes in vocal effort result in a significant impact on speaker identification (speaker ID) performance, and whisper results in the most serious loss of system performance. Given that speaker dependent whisper adaptation data is generally not available in real scenarios, these differences present a major challenge in maintaining effective speaker ID system performance.

Recently, several studies have considered compensation of whisper for neutral/whisper mismatch in speaker ID systems. In [6], an 8%–33% relative improvement in speaker ID was achieved using 5 to 15 seconds of whispered speech per speaker. In [7] and [8], compensation strategies based on frequency warping, score competition, and feature mapping were discussed for whispered utterances, based on a corpus of ten male subjects. Our study in [9] suggested that features based on short-time analysis fail to capture the most salient speaker ID information for whisper frames with low signal-to-noise ratio (SNR), based on a corpus of ten female subjects. However, no study has considered in detail the variations between whispered and neutral speech among different speakers, and their effect on neutral-trained Mel-frequency cepstral coefficient–Gaussian mixture model (MFCC-GMM) speaker ID systems. Some studies have considered silent speech interfaces (SSIs), which replace traditional acoustic microphones with sensors that capture unvocalized speech [10]–[12]. However, these SSI systems require a contact microphone, ultrasound imaging, or magnetic field sensors, which preclude practical use outside specialized applications.

The present study employs a corpus of whisper utterances from the UT-VocalEffortII corpus. Section II discusses production and acoustic characteristics of whispered speech. In Section III, two neutral trained closed-set speaker ID systems

[1]In this study, the term "neutral speech" refers to modal speech produced at rest in a quiet soundbooth.

Manuscript received November 01, 2009; revised March 05, 2010 and September 13, 2010; accepted September 25, 2010. Date of publication December 13, 2010; date of current version May 13, 2011. This work was supported in part by the Air Force Research Laboratory (AFRL) under a subcontract to RADC, Inc., under FA8750-09-C-0067, and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. Approved for public release; distribution unlimited. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nestor Becerra Yoma.

The authors are with the Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080-1407 USA (e-mail: john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2091631

based on a PLP-GMM, and an MFCC-GMM architecture are formulated. It will be shown that for whisper speaker recognition, the degradation is concentrated among a certain subset of speakers, while system performance for other speakers is consistent with that of neutral speech (e.g., the degradation in performance for speaker ID under whisper is not the same for all speakers). In Section IV, audio streams with whisper and neutral speech are each separated into two classes: unvoiced consonants, and other phonemes, based on the fact that whispered and neutral speech is similar for unvoiced consonants [2], [13].[2] The Kullback–Leibler divergence between whisper and neutral GMMs of each speaker trained with unvoiced consonants, and other phonemes, is proposed in order to assess separability, and speaker ID experimental results based on these two phoneme clusters are analyzed. It will be shown that for unvoiced consonants, the performance degradation of whisper speaker recognition is not as severe as that for other phonemes. Alternative feature extraction methods based on exponential and linear frequency scales are then proposed, to address unvoiced consonants, and speaker recognition evaluations demonstrate the methods' effectiveness in capturing speaker dependent information. Section V analyzes the acoustic properties that cause degradation in phonemes other than unvoiced consonants, and a confidence space is proposed to detect the quality of whispered speech for the speaker ID task. Finally, a speaker ID system that addresses the neutral/whisper mismatched situation is introduced in Section VI, followed by evaluations. Conclusions and a summary of this study are drawn in Section VII.

II. Whispered Speech Production and Acoustic Characteristics

In neutral speech, voiced phonemes are produced through a periodic vibration of the vocal folds, which regulates air flow into the pharynx and oral cavities. However, for whispered speech, the shape of the pharynx is adjusted such that the vocal folds do not vibrate, resulting in a continuous air stream without periodicity [14]–[18]. Changes in focal fold physiology due to functional voice disorders, trauma, or disease can cause changes to speech production which can often appear to take on whisper speech characteristics. Previous algorithms have focused on modeling and detection of such changes in speech production traits [15], [16]. However, here the focus is on a healthy vocal system. Fig. 1 shows the dramatic difference between neutral and whispered speech waveforms of the sentence "Guess the question from the answer" from the same speaker. The whispered speech waveform is lower in amplitude and lacks periodic segments.

Significant differences in the speech production process result in the following acoustic differences between neutral and whispered speech: first, there is no periodic excitation or harmonic structure in whispered speech. Second, the locations of lower frequency formants in whispered speech are shifted to higher frequencies compared to neutral speech [2]. Third, the spectral slope of whispered speech is flatter than that of neutral speech, and the duration of whispered speech is longer than that of neutral speech [1]. Fourth, the boundaries of vowel regions in the F1-F2 frequency space also differ from neutral speech [2], [19], [20]. Finally, whispered speech has much lower energy compared with neutral speech. Due to these differences, traditional neutral trained speaker ID systems degrade significantly when tested with whispered speech.

[2]Unvoiced consonants include fricatives, stops, and affricates. Non-unvoiced consonants include vowels, nasals, glides, liquids, and diphthongs, which can be provided in either a "voiced" excitation mode under neutral speech production, or an "unvoiced" excitation mode as seen under whispered speech.

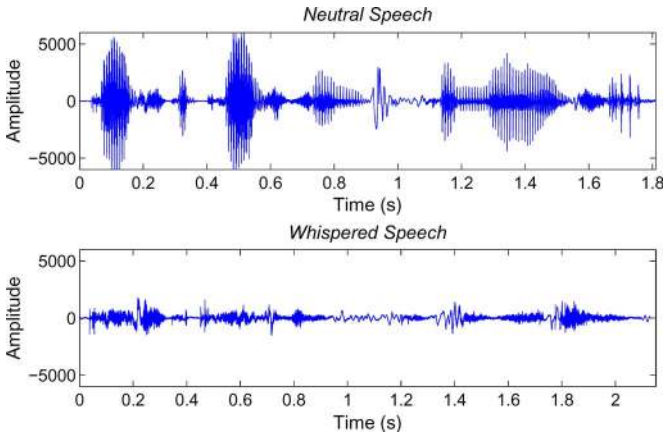


Fig. 1. Waveforms of neutral and whispered speech.

III. Baseline System Design and Performance Characteristics

A. Corpus

This study employs the UT-VocalEffort II corpus developed in [21]. The corpus consists of a total of 112 speakers, with 37 males and 75 females. Whispered and neutral speech from a subset of 28 female native speakers of American English are used in our study. The corpus consists of both read and spontaneous parts, with three sections making up the read part. In the first read section, each subject reads 40 sentences in both neutral and whisper mode. The text content of these sentences was drawn from the TIMIT database and the sentences are phonetically balanced. In the second section, two paragraphs from a local newspaper were read by each subject. For each paragraph, four whisper-islands were produced with each island including 1–2 sentences. In the third section, the paragraphs from the second section were read again. However, instead of sentences, five phrases were read in whispered mode, with each phrase consisting of 2–3 words in duration. For the spontaneous part, the data collection environmental was designed to result in a natural conversation [21]. The spontaneous part consists of two sections. In the first section, a list of randomly organized pieces of Key information, including names, addresses, phone numbers, and credit card numbers were provided to each subject. Key information was randomly chosen to be spoken either with whispered or neutral mode. In the second section, each subject was asked ten questions, such as "name three of your favorite movies," "what do you think of the weather in Dallas." Each subject was free to choose to answer three of the topic questions with whisper and the remaining seven with neutral

TABLE I
CLOSED-SET SPEAKER RECOGNITION PERFORMANCE FOR MFCC AND PLP BASED SYSTEMS

Speech Mode		Accuracy (%)		
Training	Testing	static MFCCs	PLPs	static+delta MFCCs
Neutral	Neutral	99.10	96.97	99.27
Neutral	Whisper	79.29	43.91	60.04
Whisper	Whisper	94.41	93.01	94.78
Whisper	Neutral	9.45	8.91	8.18

speech. The answers were limited to 1–2 sentences. The whispered and neutral streams of all subjects were manually separated to constitute the whisper and neutral corpora. From [21], we also note that all recordings include a 1-kHz 75 dB pure-tone calibration sequence to provide ground-truth on vocal effort for all speakers and sections. Speech data was digitized using a sample frequency of 16 kHz, with 16 bits per sample. Speech from all speakers was windowed using a Hamming window of 32 ms, with an overlap of 16 ms for the entire speech processing phase in this study.

B. Baseline System Design

In this section, closed-set speaker recognition experiments are conducted using read whispered and neutral speech in train/test matched/mismatched conditions. Three different commonly used features are considered for our baseline system: 19-dimensional static MFCCs, 38 dimensional static+delta MFCCs, and 12-dimensional static perceptual linear predictive coefficients (PLPs) [22]. In the train/test configuration for the neutral/neutral scenario, an average of 50 read neutral train utterances and 20 read neutral test utterances are used per speaker. For the train/test of the neutral/whisper scenario, an average of 70 neutral read utterances from each speaker are used for training and an average of 34 whispered read utterances of each speaker are used for testing. For the whisper/whisper scenario, an average of 20/14 whispered read utterances from each speaker are used for train/test. Finally, for the whisper/neutral scenario, on average a set of 20 whispered read utterances from each speaker are used for training and 20 neutral read utterances are used for test. The differences in each of these scenarios for train/test are due to the available native speech material from UT-VocalEffort II.

In our baseline system, silence parts of whisper and neutral speech are first removed using a dynamic energy threshold that depends on the SNR of each particular sentence block sequence. Next, a 64-mixture universal background model (UBM) is constructed using the Expectation–Maximization (EM) algorithm with features extracted from all available training data. A Gaussian mixture model (GMM) for each speaker is obtained afterwards using maximum *a posteriori* (MAP) adaptation with whispered and neutral speech, respectively. The results of closed-set speaker ID are listed in Table I.

From Table I, it can be seen that for the matched train/test whisper condition, the performance of MFCC-GMM and PLP-GMM systems are comparable to the traditional neutral/neutral matched train/test condition. This illustrates that whispered speech contains sufficient speaker information for effective speaker ID. However, due to the lack of excitation, and vocal tract dependent differences between whispered and neutral speech, a portion of the speaker dependent information carried in neutral speech is lost or distorted in whispered speech (e.g., 94.41% versus 99.10%). Based on the four mismatched/matched scenarios, a number of key observations are clear from the results. First of all, for the train/test in neutral/whisper, results from Table I suggest that static MFCCs generally outperform static PLPs for our speaker ID task. It is noted as well, that the static MFCC system outperforms the static + delta MFCCs system for neutral/whisper train/test scenario. This is because the duration of whispered speech is generally longer than that of neutral speech. The results also suggest that there is speaker dependent structure contained in neutral speech that is also presented in whispered speech, but sufficient differences exist that cause a significant loss in closed-set speaker ID performance (e.g., 99.10% versus 79.29%). Alternatively, for the train/test in whisper/neutral, there is a profound loss in speaker ID performance (e.g., 9.45% versus 99.10%). This suggests that there is speaker-dependent structure within neutral speech that can be used for training to capture specific speaker structure in whisper. However, the speaker dependent structure of whispered speech is so profoundly different from the structure of neutral speech, that training with whisper for neutral speaker ID becomes infeasible. Based on the performance for the train/test neutral/whisper condition, we choose the static MFCC-GMM as our baseline system.

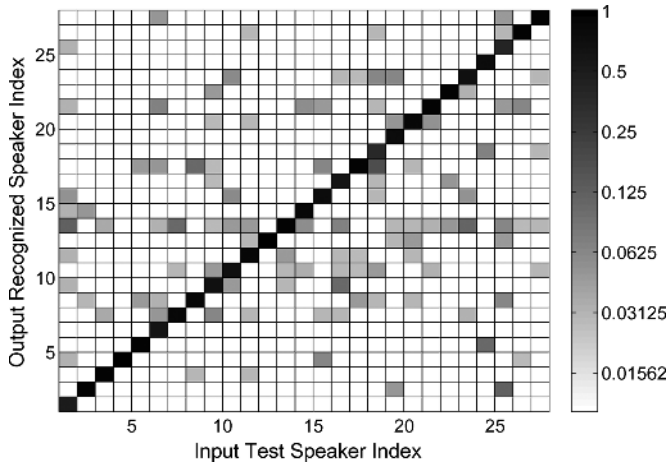


Fig. 2. Confusion matrix for neutral trained MFCC-GMM system when tested on whispered utterances. The matrix entries are displayed graphically by a logarithmic gray scale according to the legend at right.

C. Speaker ID Performance Variations Among Different Speakers

This section discusses the variation in speaker ID performance for whispered speech among different speakers. A confusion matrix for the static MFCC-GMM system is shown for the neutral/whisper train/test scenario in Fig. 2. The gray scale key to the right indicates that darker shades denote a higher probability that the input speaker is recognized as the corresponding correct speaker. Thus, darker diagonal entries indicate stronger system performance. The number along the x and y axis is the closed-set speaker index. From Fig. 2, it can be observed that the accuracy varies significantly across speakers. For example, for Speakers 4 and 12, an accuracy of 100.0%

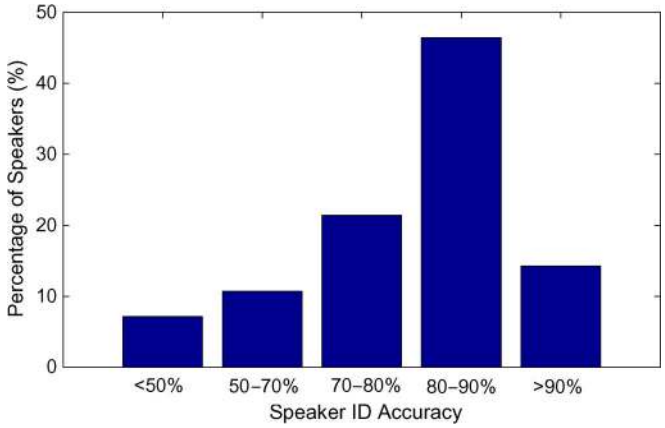


Fig. 3. Distribution of accuracy among speakers.

is achieved when models are trained with neutral speech and tested with whispered speech. However, for Speaker 18, only 46.0% of the whispered utterances are correctly identified. There are also 19 speakers from the total of 28 speakers for which system accuracy is between 70%–90%. These results illustrate an interesting property of whispered speech: some speakers maintain sufficient speaker dependent structure under whisper condition, so that no additional processing or system changes are needed when neutral speech models are employed, while others fail completely. This suggests a new approach to whisper speech based speaker ID, where "consistent" or "good" speakers are identified with no additional processing, and others are routed to alternative compensation methods.

Fig. 3 summarizes the distribution of closed-set speaker recognition accuracy across the speaker set using neutral trained models and whisper test material. The y axis represents the percentage of speakers whose overall closed-set speaker ID accuracy(%) falls into the accuracy range listed along the x axis. From Fig. 3, it can be seen that whispered utterances from 14.3% of the input speakers can be used for speaker recognition with a >90% accuracy, while 10.7% of the speakers can only be recognized correctly with whispered speech 50%–70% of the time. An assessment of speaker information, such as their age, height, weight, years of formal education, and place of birth, was also considered to determine if such factors might influence or contribute to speaker ID performance degradation between neutral/whisper. However, no factors were found to be statistically correlated with sustained whisper performance using a neutral-trained speaker ID system. The speaker recognition results here suggest that the similarity of speaker information between neutral speech and whispered speech depends on how the speaker produces whisper. We will refer to those whisper utterances that contain speaker dependent structure sufficiently similar to that seen in neutral utterances to be correctly recognized by a neutral-trained system as "high-performance whisper" (HPW) in the remainder of this paper. As such, those whisper utterances that are not valid for speaker ID with a neutral trained system will be referred to as "low-performance whisper" (LPW). It should be noted that LPW does contain speaker information. However, the dissimilarity of speaker information in these utterances to speaker information in neutral speech results in low system performance.

IV. Phoneme Analysis for Speaker ID With Whispered Speech

A. Baseline Phoneme Analysis

In [2] and [10], it was shown that the acoustic characteristics of vowels and voiced consonants differ significantly between whispered and neutral speech, while the spectral properties of unvoiced consonants are relatively similar between whispered and neutral speech. This section addresses the question: Is the degradation in speaker ID for the neutral/whisper mismatched scenario dependent on those two broad phoneme classes?

This phase employs only the read portion of the UT-Vocal-Effort II corpus. Unvoiced consonants in both neutral and whispered speech are first separated from all other phonemes. For simplicity, this section will refer to phonemes other than unvoiced consonants, such as vowels, liquids, glides and diphthongs, as non-unvoiced consonants.[3] It is noted that, due to the absence of periodic excitation, the voiced consonants in whispered speech are similar to the unvoiced consonants. This implies that voiced/unvoiced phoneme pairs, such as (z, s), (zh, sh), (t, d), (dzh, tsh), etc., are not separated. In order to determine those frames that belong to unvoiced consonants, three measurements are made for each frame i: the total energy in each of the frequency bands: 100–4000 Hz, 4000–8000 Hz, and 100–8000 Hz. These measurements are denoted as $f_{i,l}$, $f_{i,h}$, and $f_{i,e}$ (e.g., low "l" freq, high "h" freq, and everything "e" in the frequency range with "i" as the frame). Next, the ratio of $f_{i,l}$ to $f_{i,e}$ is calculated and denoted as $Rn(i)$, where i is the frame index. The relative symmetric entropy in the frequency domain is also calculated as shown in (1) in order to compare the spectral structure of neighboring frames in the high frequency domain

$$En(i) = -P_{h,i-1}log(P_{h,i}) - P_{h,i}log(P_{h,i-1}) \qquad (1)$$

where $P_{h,i} = f_{i,h}/f_{i,e}$. This term is obtained because most of the spectral energy of unvoiced consonants is concentrated in the higher frequency domain, while the energy of most vowels is concentrated in lower frequencies. A threshold can be set for both Rn and En to separate out unvoiced consonant frames. Experimental results demonstrate the effectiveness of this method in detecting unvoiced consonants. Therefore, we obtain an unvoiced consonant neutral/whisper corpus and a non-unvoiced consonant neutral/whisper corpus for analysis by using this procedure. Further discussion is included and associated with Fig. 12 in Section VI. Next, 19-dimensional static MFCC vectors are extracted as features for training and testing the speaker ID systems.

Four 64-mixture GMMs are trained for each speaker using the following automatic detected data sets: neutral-unvoiced consonants (ne-uc), neutral-non-unvoiced consonants (ne-nuc), whisper-unvoiced consonants (wh-uc), and whisper-non-unvoiced consonants (wh-nuc). The Kullback–Leibler divergence [24], which provides a measure of the distance between two

[3]We note here, that these phonemes are normally voiced but under the whispered speech, there is no voiced excitation, hence we use the label "non-unvoiced."

probability distributions, is employed here to assess the distance between whisper and neutral GMM models. Each GMM is represented in as

$$f(x) = \sum_a \pi_a N(x; \mu_a, \Sigma_a) \tag{2}$$

where π_a is the prior probability of each mixture, and $N(x; \mu_a, \Sigma_a)$ is a multi-dimensional (19-D) Gaussian with an observation vector x, mean vector μ_a, and covariance matrix Σ_a. Here, only a diagonal covariance matrix is considered. Next, all mixtures of each GMM are fused into one mean vector and variance vector as shown as

$$\mu_{\hat{f}} = \sum_a \pi_a \mu_a \tag{3a}$$

$$\Sigma_{\hat{f}} = \sum_a \pi_a \left(\Sigma_a + (\mu_a - \mu_{\hat{f}})(\mu_a - \mu_{\hat{f}})^T \right). \tag{3b}$$

Next, the KL divergence between GMM $f(x)$ and $g(x)$ can be obtained through (4)

$$D(\hat{f}||\hat{g}) = 0.5 \left[log\frac{\Sigma_{\hat{g}}}{\Sigma_{\hat{f}}} + Tr\left[\Sigma_{\hat{g}}^{-1}\Sigma_{\hat{f}}\right] - dim + (\mu_{\hat{f}} - \mu_{\hat{g}})^T \Sigma_{\hat{g}}^{-1} (\mu_{\hat{f}} - \mu_{\hat{g}}) \right] \tag{4}$$

where dim is the dimension size. A symmetric KL divergence between $f(x)$ and $g(x)$ is thus obtained by (5). If two GMMs are similar, the KL divergence between them will be small (but not necessarily zero). Alternatively, if two GMMs are quite different, their KL divergence will be large.

$$D(\hat{f}, \hat{g}) = D(\hat{f}||\hat{g}) + D(\hat{g}||\hat{f}). \tag{5}$$

The KL divergence is first calculated between the ne-uc GMM and wh-uc GMM for each speaker. The same procedure is conducted for the ne-nuc GMM and wh-nuc GMM for each speaker. For unvoiced and non-unvoiced consonants, the KL divergences are also normalized respectively with the means for comparison. Fig. 4(a) shows the result of the divergence between the ne-uc GMMs and wh-uc GMMs corresponding to each speaker, and Fig. 4(b) presents the same results for the ne-nuc GMMs and wh-nuc GMMs.

From Fig. 4, it can be seen that the KL divergence of the unvoiced consonant GMMs between whispered and neutral speech is much smaller compared to that for the non-unvoiced consonant GMMs. For example, the KL divergence ranges from 0.05–0.34 among all 28 speakers for unvoiced consonant GMMs, with an average value of 0.16, while for the non-unvoiced consonants, the KL divergences range from 0.3–1.3, with an average value of 0.64. This result confirms the observation that unvoiced consonants do not vary as much as other phonemes between whispered and neutral speech. It is also noted that, compared with non-unvoiced consonants, the variance of the KL divergence between whispered and neutral speech among different speakers is generally much smaller for unvoiced consonants. The normalized KL divergence ranges from -0.1 to $+0.18$ among all 28 speakers for unvoiced consonant GMMs. However, for those GMMs constructed with other phonemes (non-unvoiced consonants), the normalized KL divergence ranges from -0.33 to $+0.67$ and has a larger variance across the speaker pool. Therefore, we conclude that the loss in speaker ID performance due to whispered test data is primarily due to changes in the acoustic properties of non-unvoiced consonants.

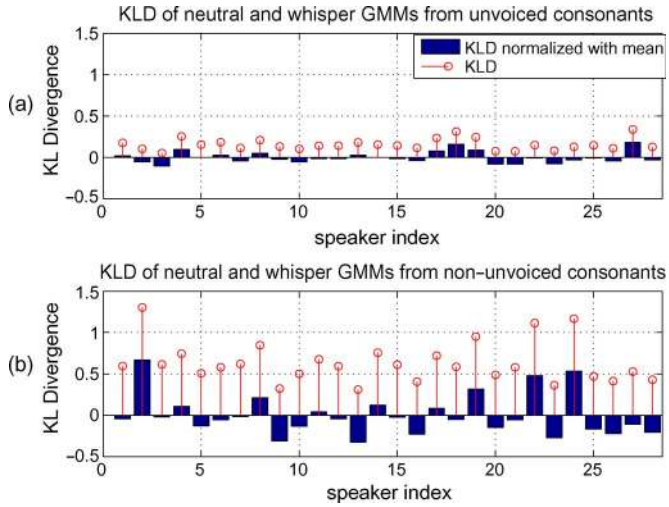


Fig. 4. KL divergence of (a) unvoiced, and (b) non-unvoiced consonant GMMs between whispered and neutral speech.

B. Speaker Recognition Based on Unvoiced/Non-Unvoiced Consonant GMMs

In this section, speaker recognition experiments are conducted based on ne-uc GMMs and ne-nuc GMMs. For unvoiced consonant GMMs, the test data is drawn from the read whispered and neutral unvoiced consonant sections for all speakers, and for the non-unvoiced consonant GMMs, the test data is obtained from the read whispered and neutral non-unvoiced consonant portions from all speakers. Table II summarizes the closed-set speaker ID results. It can be seen that for the non-unvoiced consonant part, the mismatch between whisper and neutral produces a significant performance degradation, with an absolute 40.8% reduction in speaker ID performance. However, for the unvoiced consonant portion, whispered speech causes only a 7.27% degradation. It is noted that the performance of GMMs trained with unvoiced consonants is not comparable to those trained with non-unvoiced consonants, when testing is performed with neutral speech. This is mainly caused by two reasons: first, there is limited formant structure contained in unvoiced consonant speech, except for those formant structures stemming from neighboring vowels. Thus, the potential speaker dependent information contained in unvoiced consonant speech is limited and cannot convey sufficient speaker dependent structure compared to that from other phonemes (mostly vowels). Second, since most of the spectral information from unvoiced consonant speech is contained in high frequencies, feature vectors that emphasize more low frequency structure, such as MFCCs, cannot completely capture the necessary spectral structure in the high frequency domain. In order to capture as much speaker-dependent information from unvoiced

TABLE II
RECOGNITION RESULT FOR UNVOICED/NON-UNVOICED CONSONANT

Speech Mode		Phoneme	Accuracy
Training	Testing		
Neutral	Neutral	non-unvoiced consonant	98.14%
Neutral	Whisper	non-unvoiced consonant	57.34%
Neutral	Neutral	unvoiced consonant	73.76%
Neutral	Whisper	unvoiced consonant	66.49%

consonant speech as possible, a new feature extraction method is developed, in place of MFCCs, in the next section.

C. Modified LFCC and EFCC

Since there are differences in the spectral energy distribution between unvoiced consonants and other phonemes, feature vectors based on alternative scales are expected to perform better than applying a single scale for all phonemes. Also, based on analysis from Section IV-B, feature extraction methods that emphasize low frequency components more than high frequency components, such as MFCCs, are not effective for unvoiced-consonant phonemes. This section discusses feature extraction methods based on both linear and exponential frequency scales similar to that proposed for speech under stress in [25]. Different from our study in [7], [8], here we propose to apply both scales for feature extraction of only unvoiced consonants.

The general form of the exponential mapping function first proposed in [25] is employed in this study

$$y = c \times \left(10^{\frac{f}{k}} - 1\right), \quad 0 \le f \le 8000 \text{ Hz}. \tag{6}$$

The values of c and k are obtained by solving the following two equations:

$$c \times (10^{8000/k} - 1) = 2595 \times log\left(1 + \frac{8000}{700}\right) \tag{7a}$$

$$\{c,k\} = min\left\{\left|(10^{4000/k} - 1) - \frac{4000}{k^2} \times ln10 \times c \times 10^{\frac{4000}{k}}\right|\right\}. \tag{7b}$$

(7a) is obtained by requiring that the exponential and Mel-scale warping functions be equal at the Nyquist frequency. (7b) is obtained by minimizing the absolute value of the partial derivatives of (6), with respect to c and k when $f = 4000$ Hz. Here, $f = 4000$ Hz is chosen based on our experiment in [7]. After solving (7a) and (7b), we obtain $c = 6375$ and $k = 50000$. Hence, the exponential scale function used here will be

$$y = 6375 \times (10^{f/50000} - 1). \tag{8}$$

Fig. 5 illustrates the three mapping functions: 1) the original mel-scale, 2) the explored exponential-scale, and 3) the linear-scale. The computation of the cepstral coefficients based on the exponential-scale and linear-scale mapping functions is similar to MFCCs. A set of 26 triangular bandpass filters are placed according to the corresponding mapping function. Next, the cosine transform is applied to the log energies obtained from the filter banks with cepstral liftering applied afterwards. For simplicity, the cepstral coefficients obtained through the exponential mapping function will be refereed to as EFCCs, and

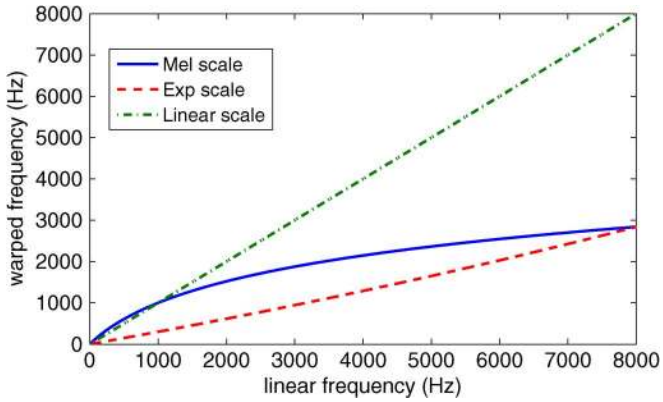


Fig. 5. Mel, exponential, and linear scale warping function.

TABLE III
CLOSED-SET SPEAKER RECOGNITION RESULTS FOR UNVOICED CONSONANT SPEECH

Speech Mode		Feature Vector	Accuracy
Training	Testing		
Neutral	Whisper	MFCC	67.01%
Neutral	Whisper	LFCC	71.90%
Neutral	Whisper	EFCC	70.55%

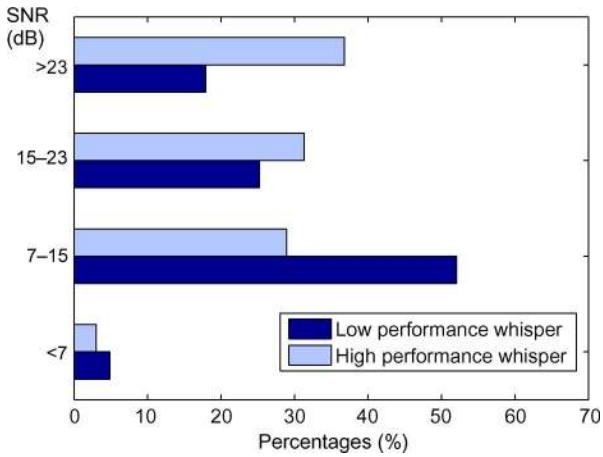


Fig. 6. Comparison of SNR for high- versus low-performance whispered speech for speaker ID.

those from the linear mapping function will be referred to as LFCCs.

Speaker recognition experiments are conducted in order to compare performance for MFCC, LFCC, and EFCC features for unvoiced consonants. GMMs for each speaker are trained with 19-dimensional static MFCC, LFCC, and EFCC features extracted from unvoiced consonants. Because there is very little spectral energy in the lower frequencies of most unvoiced consonants, the starting frequency of MFCC, LFCC, and EFCCs is moved above 0 Hz. Our experiments show that an upper frequency value of 300 Hz results in the best performance. Table III summarizes the speaker recognition results based on MFCC, LFCC, and EFCCs. The results show that the best performance is achieved with LFCCs, with a 4.89% absolute improvement, compared to the system based on MFCCs. It is noted that while this may not be large, it is obtained with little change in the computational requirements or training paradigms. Performance of the system based on EFCCs also

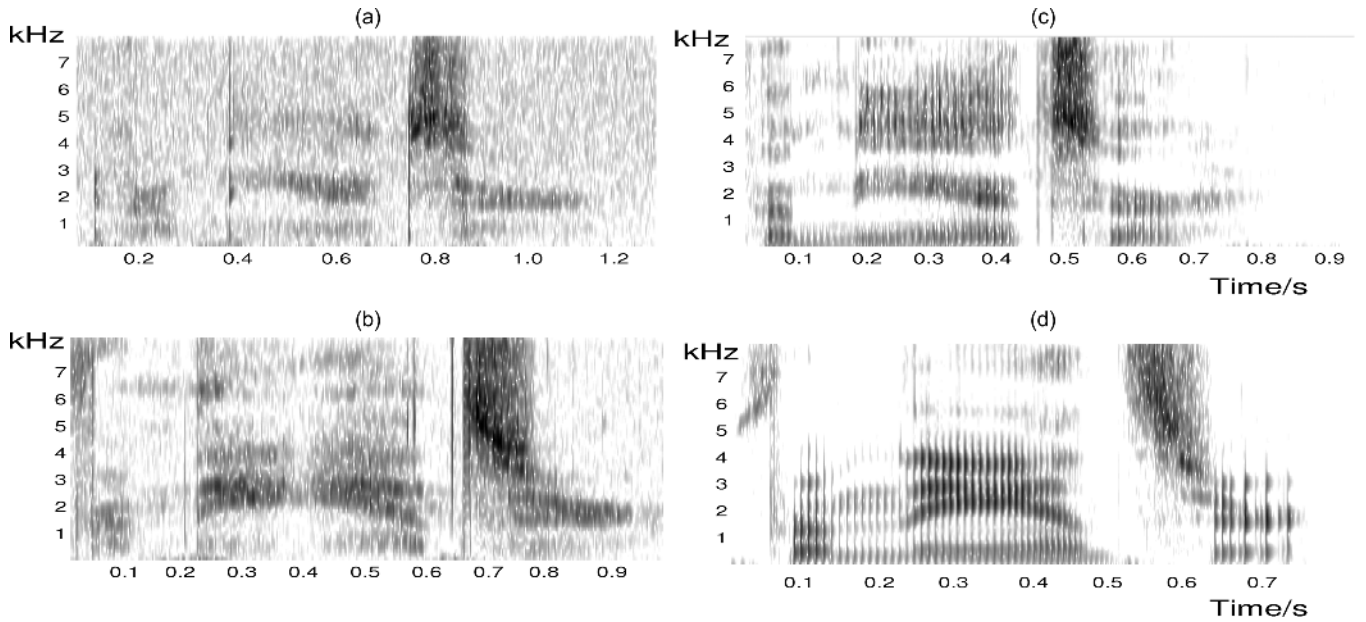


Fig. 7. Phrase "from the answer" from two speakers in both neutral and whispered speech mode: (a) lower SNR whisper from speaker I, (b) higher SNR whisper from speaker II, (c) neutral speech from speaker I, (d) neutral speech from speaker II.

improves compared with MFCCs. This result confirms that a frequency scale that emphasizes higher frequency components is more effective for feature extraction of unvoiced consonants in speaker recognition systems.

V. Analysis of Degradation in Non-Unvoiced Consonant for Speaker ID

Section IV-B showed that the mismatch between whispered and neutral speech, which causes performance degradation in neutral trained speaker ID, comes mainly from non-unvoiced consonants. This section considers the acoustic differences between whisper and neutral speech among those phonemes, and its effect on the performance of an MFCC-GMM system.

A. *Effect of SNR*

This section analyzes the effect of SNR on the performance of a speaker ID system for the neutral/whisper mismatched condition. As noted earlier, whispered speech is divided into two sets: high- and low-performance whisper. The SNRs of high- and low-performance whisper are calculated based on the average energy within silence versus the average energy of the speech part for each whisper utterance, and are further classified into four SNR sets: $<$7 dB, 7–15 dB, 15–23 dB, and $>$23 dB. The proportion of utterances in each of the SNR ranges among all utterances is calculated separately for high- and low-performance whisper. The results are shown in Fig. 6.

From Fig. 6, it can be seen that a large percentage of the high-performance whisper utterances are of larger SNR than that of low-performance whisper. For example, the percentage of high-performance whisper with SNR above 23 dB is 37%, which is 20% higher than that of the low-performance whisper speakers. Meanwhile, 53% of the low-performance whisper has an SNR between 7–15 dB, while only 29% of high-performance whisper is in this area. This result suggests that the estimated SNR may be used as a feature to differentiate high- and low-performance whispered speech.

SNR may provide a clue to speaker ID performance due to the inherent limitations of features based on short-time analysis. We note that all audio files are noise-free, so SNR here is related to the balance between silence and speech signal energy. Feature vectors, such as MFCC and LFCC, attempt to represent the spectral energy/power envelope along the continuous frequency domain (for example, from 100–8000 Hz). Hence, when part of the spectral information is lost because of low vocal effort, the neutral-trained model will fail to capture some speaker-dependent components for the mismatched whispered test data. Fig. 7 shows spectrograms of the phrase "from the answer" from two distinct speakers in both neutral and whispered speech modes. One speaker [Fig. 7(a)] demonstrates a lower SNR whisper with 8.2 dB (note: the corresponding neutral speech has an SNR of 25.7 dB), while the second speaker [Fig. 7(b)] displays a SNR of 20.0 dB (with the corresponding neutral speech having an SNR of 27.9 dB). As seen in Fig. 7, for whisper with high SNR, despite the differences that exist with respect to neutral speech including formant shifting and the absence of excitation, the overall spectral structure is generally preserved. Therefore, feature vectors based on MFCCs still possess similar structure between neutral and whispered speech. However, for whisper with low SNR, due to their low volume caused by the way the speaker produces whisper, a portion of the formant structure is lost or reduced relative to the background silence floor, or in some cases completely missing, even though the actual background silence/noise level is very low. Hence, it is more likely that feature extraction methods based on short-time analysis will cause neutral trained systems to degrade significantly when tested with lower SNR whispered speech.

It should be noted that there remain a number of subjects with relatively low SNR whisper that are still recognized correctly using neutral speaker ID models. For example, even when the SNR is below 7 dB, among all the whisper utterances belonging to both high- and low-performance groups in this area, 40% of the utterances still provide correct speaker recognition

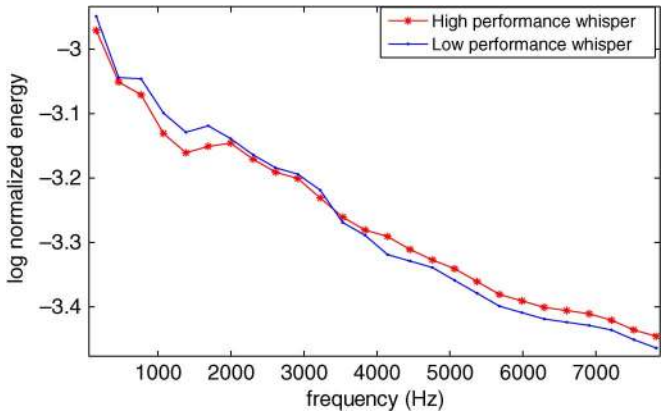


Fig. 8. Comparison of normalized mean of UBMs for high/low-performance whisper.

performance. When the SNR is between 7–15 dB, 35% of the whisper speakers are correctly recognized. Therefore, the SNR is not the only property that dictates the quality of whispered speech performance for speaker recognition. The next section explores other acoustic properties that impact speaker ID system performance.

B. Effect of Other Acoustic Properties for Whisper

It has been shown that the SNR for whispered utterances can be employed to identify high- versus low-performance whispered speech for speaker ID. In this section, we further explore the differences between these two kinds of whispered speech, in order to identify acoustic properties that lead to improved performance levels.

In order to compare differences in spectral structure between high- and low-performance whisper, two UBMs are trained using the log energy output from a 26-band linear scale filterbank. The means of each Gaussian mixture of the UBM are fused using (3a), and normalized with respect to the total power. The log of the mixture mean is plotted for high- and low-performance whisper UBMs in Fig. 8.

From Fig. 8, it can be seen that the distribution of spectral energy of high- and low-performance whisper differs mainly in two ways: first, the spectral slope for low-performance whisper is greater (i.e., steeper slope) than that of high-performance whisper. This result suggests that the spectral slope is related to the performance of the neutral trained speaker ID system on whispered speech. This phenomenon is related to the fact that whispered and neutral speech share similar structure in the higher frequencies [5], [8]. Therefore, observations that contain more spectral information in higher frequencies should have a higher probability of correct speaker recognition using a neutral-trained speaker ID system. We can also observe from Fig. 8 that high-performance whisper has a clear reduction of spectral energy between 1000–2000 Hz compared to low-performance whisper. Before suggesting an explanation, it is noted that formant shifts between whispered and neutral speech occurs primarily in the F1, F2 region, and there are almost no shifts towards high frequency in the range of F3 and F4 [2], [4]. Therefore, whisper has a lower chance of being correctly recognized for speaker ID because more information exists in the F1 and F2 area, and less in F3 and F4 area. This suggests an explanation for why low-performance whisper generally has a higher percentage of spectral energy between 1000–2000 Hz than high-performance whisper.

TABLE IV
COMPARISON OF SPECTRAL TILT FRAME DISTRIBUTION BETWEEN HIGH- AND LOW-PERFORMANCE WHISPER

Spectral tilt (dB/octave)	Whisper	
	Low performance	High performance
< -2	75.6%	52.6%
-2 to 1.5	23.58%	40.6%
-1.5 to 1.0	0.0%	5.0%
> -1.0	0.8%	1.8%

TABLE V
COMPARISON OF $ratio^{f}_{1-2vs1-8}$: RATIO ENERGY IN 1–2 kHz VERSUS 1–8 kHz

$ratio^{f}_{1-2vs1-8}$	Whisper	
	Low performance	High performance
< 0.3	0.8%	28.6%
$0.3 - 0.4$	41.5%	47.6%
$0.4 - 0.5$	45.5%	19.3%
> 0.5	12.2%	4.5%

Based on observations from Fig. 8, two acoustic properties are suggested in order to further separate high/low-performance whisper. First, spectral tilt is used in order to measure the energy of the high-frequency components. The method developed by Hansen [26] is applied to calculate the spectral tilt for high- and low-performance whisper utterances. Table IV lists the frame distribution of spectral tilt among the two performance classes of whisper, respectively. It is observed that 75.6% of the lower performance whisper frames has a spectral tilt that is lower than -2 dB/octave compared with 52.6% for high-performance whisper speakers. 40.6% of the high-performance whisper frames for all speakers have a spectral tilt between -2 to -1.5 dB/octave compared to 23.58% of the low-performance whisper. For a spectral tilt above -1.5 dB/octave, only 0.8% of the low-performance whisper frames has a spectral tilt in this area compared with 6.8% of high-performance whisper. The results in Table IV confirm our observations from Fig. 8.

Next, in order to measure the percentage of spectral energy present in the 1000–2000 Hz band (based on the energy gap seen in Fig. 8), we calculate the ratio of the energy in 1000–2000 Hz to the energy in 1000–8000 Hz, which will be referred to as $ratio^{f}_{1-2vs1-8}$. The distribution of $ratio^{f}_{1-2vs1-8}$ among high/low-performance whisper speakers is summarized in Table V. This result also confirms the observation from Fig. 8, where whisper frames with more energy in 1000–2000 Hz performs more poorly for speaker recognition. For example, only 0.8% of the low-performance whisper frames has a $ratio^{f}_{1-2vs1-8}$ lower than 0.3. However, 28.6% of the high-performance whisper has a $ratio^{f}_{1-2vs1-8}$ in this area. Meanwhile, there is 57.7% of low-performance frames with $ratio^{f}_{1-2vs1-8}$ above 0.4 compared with 23.8% of high-performance whisper frames. This suggests that the spectral slope and $ratio^{f}_{1-2vs1-8}$ could be used as complimentary indicators of high versus low-performance whisper based speakers for speaker recognition.

C. Confidence Space

In this section, we establish a process for identifying speakers whose whispered speech is effective for speaker ID. It is noted that most compensation strategies developed so far for whispered speech attempt to conduct compensation directly on all whispered speech [7], [8]. Because this study has shown that a portion of the whispered speech may be used directly for speaker recognition, compensation is not necessary for all whispered speech, and in fact may even reduce system performance if processing is applied to high-performance whisper. Sections V-A and V-B considered the effects of SNR, spectral tilt and $ratio^f_{1-2vs1-8}$ on performance degradation of a MFCC-GMM speaker ID system for whispered non-unvoiced consonants. Based on the results obtained from these two sections, we propose a two-dimensional feature space to measure the quality of whispered speech with the goal of achieving improved speaker ID performance.

First, since the $ratio^f_{1-2vs1-8}$ and spectral tilt both represent the distribution of spectral energy along the frequency axis, they are combined in order to constitute an SR parameter (spectral representative) as follows:

$$SR = \frac{1}{ratio^f_{1-2vs1-8} \times (1 - e^{0.2 \times tilt})}. \tag{9}$$

This SR parameter is found experimentally. The nonlinear term, with $tilt$, is introduced so the value of SR will increase faster with a larger value of spectral tilt (such as > -1.5 dB/octave) than if we employ a linear term of spectral tilt, such as $-1/(ratio^f_{1-2vs1-8} \times tilt)$. In this case, if a whisper utterance has a spectral tilt above -1.5 dB/octave, there is more chance that it will have a larger value of SR, even though it has a relatively high $ratio^f_{1-2vs1-8}$ (such as 0.5) that will otherwise produce a smaller value for SR. It is noted that the range of values of SR is a number greater than zero and generally less than 40, based on the speech features extracted in our study.

Next, the SR and SNR terms can be employed to constitute a two dimensional space, which will be referred to as the confidence space for assessing the quality of the input whisper data. For each test whisper utterance, we obtain SR and SNR. The values of these two parameters provide a position in the confidence space, which represents whether this is believed to be a high- or low-performance whisper for speaker recognition. According to this position, we can either decide to perform further compensation, or in a dialog system scenario the speaker can be asked to repeat the previous whispered speech utterance again, which is expected to improve the quality of whisper. In Fig. 9, the position for both high- and low-performance whisper in the confidence space is shown. It can be seen that most low-performance whisper entries converge towards the lower left corner of the confidence space, which is indicated with a square in Fig. 9. Alternatively, the high-performance whisper utterances occupy a region with relatively high SNR, and for those entries with lower SNR, their value of SR tends to have higher values. More sophisticated methods, such as support vector machines or neural networks, can also be applied to search for a more optimized boundary.

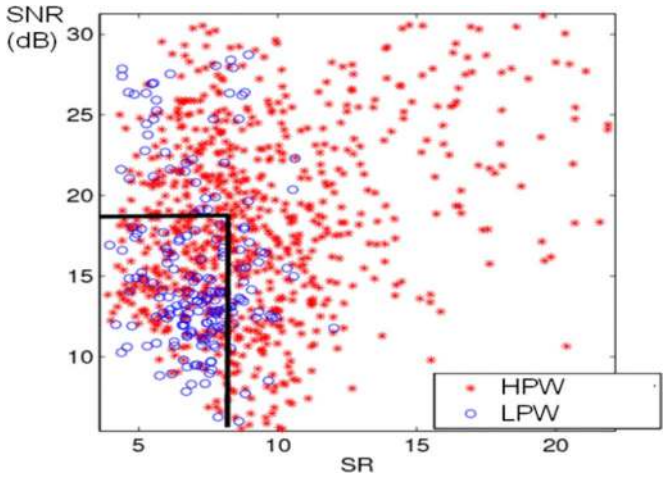


Fig. 9. Distribution of confidence space for high/low-performance whisper.

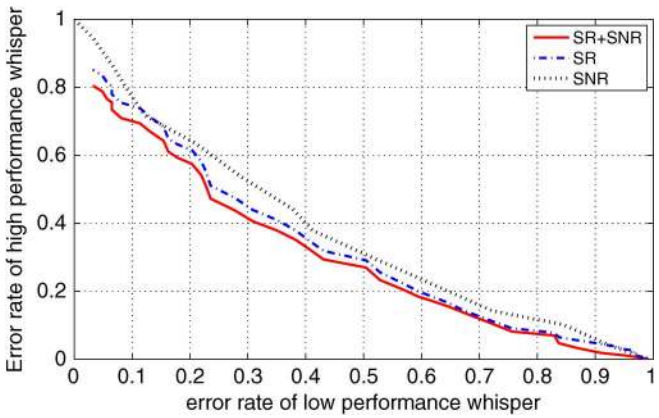


Fig. 10. Distribution of error rate for high/low-performance whisper.

By moving a hard threshold along the two SNR and SR axes, we can obtain a set of error rates shown in Fig. 10. The error rate here is calculated as

$$Error\ \ Rate = \left(1 - \frac{S}{N}\right) \times 100\% \tag{10}$$

where N is the total number of utterances of either high- or low-performance whisper, and S is the number of utterances that are classified as contributing to correct speaker recognition with the confidence space for each type of whisper. We can see from Fig. 10 that there is a tradeoff between detecting high- and low-performance whisper. For example, when 30% of low-performance whisper is misclassified as high performance using the proposed hard threshold from the confidence space, 60% of the high-performance whisper will be correctly classified. However, when 90% of the low-performance whisper can be correctly identified, only about 30% of the high-performance whisper will be considered "good" according to the hard decision threshold. However, we can also observe from Fig. 9 that even though some of the low-performance whisper is sitting outside the hard threshold, most are near the boundary, and thus the position of the confidence space can still provide valuable information as to the quality of the whisper for speaker ID. On the other hand, although part of the high-performance whisper

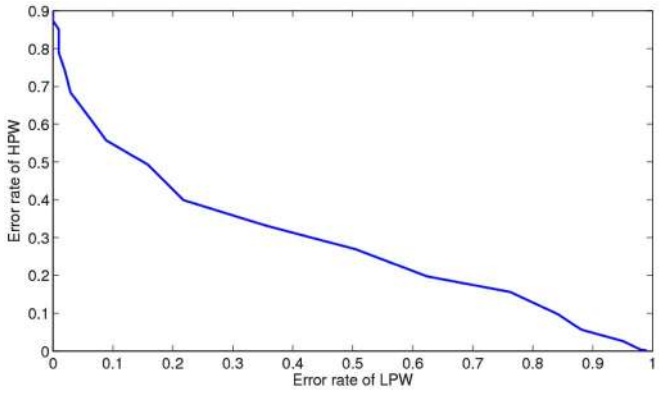


Fig. 11. Distribution of error rate for high-performance whisper (HPW) versus low-performance whisper (LPW) by fixing the SNR boundary, and varying the spectral representative (SR).

with high SNR and SR can be identified clearly as high performance in the confidence space, other parts of the high-performance whisper fall into the boundary of low-performance whisper. Considering that this part of high-performance whisper has similar acoustic properties with low-performance whisper, it is reasonable to assume that further compensation will not harm that part of the data for whisper-based speaker ID.

For comparison, we also plot the curve of error rates when only setting a threshold for SNR or SR in Fig. 10. As seen from Fig. 10, the performance using SNR combined with SR outperforms the case when either SNR or SR is used alone. This result confirms our conclusion in Sections V-A and V-B that SNR should be combined with other acoustic features for effective detection of high/low-performance whisper in speaker ID.

Finally, in order to show the effectiveness of SR in detecting high/low-performance whisper, especially for those with low SNR, we also calculate and plot the error rates using only the whisper utterances with an SNR below 20 dB while moving the boundary of SR from 2 to 15 in Fig. 11. As we can see in Fig. 11, the SR parameter can be used efficiently to detect low-performance whisper from high-performance whisper even if both have relatively low SNR values. For example, when 80% of the low-performance whisper with low SNRs are correctly identified, almost 60% of the high-performance whisper will achieve correct speaker ID performance. This proves that the SR parameter is a necessary complementary trait to SNR in order to assess the quality of whispered speech for speaker ID.

VI. Speaker Identification System for Whispered Speech and Recognition Results

A. System Description

This section proposes the overall system for speaker ID of whispered speech shown in Fig. 12, based on results obtained from Sections III–V. As discussed in Section IV-C, for unvoiced consonants, LFCC- and EFCC-based features outperformed an MFCC-based system by capturing more speaker specific information contained in high frequencies. Thus, unvoiced consonants are separated from other phonemes and processed with LFCC or EFCC feature extraction to enhance overall performance. The likelihood scores averaged over all frames in each utterance from LFCC or EFCC-GMM and MFCC-GMM-based speaker ID systems are combined together as shown in (11). Only the complete set of utterances from the neutral speech mode are used to train the MFCC-GMM speaker ID system. Next, unvoiced consonants will be separated from other phonemes within the input audio streams and employed to train an LFCC-GMM or EFCC-GMM system. For testing, unvoiced consonants are separated from other phonemes first:

$$S = \alpha S_{\mathrm{MFCC}} + (1-\alpha) S_{\mathrm{LFCC/EFCC}}. \tag{11}$$

In order to show the effect of α from (11) on both low/high-performance whisper for speaker ID, a closed-set speaker recognition experiment is conducted by adjusting the value of α from 0.1 to 1.0. The LFCC feature vectors are used here; however, the same result should be expected from EFCC-based features. When α is 1.0, the system revert back to the single MFCC-GMM baseline system. The experimental results are shown in Fig. 13. It is observed that for low-performance whisper utterances, the accuracy increases from 0% to 60% when α decreases from 1.0 to 0.1, while the accuracy of high-performance whisper drops from 100% to 80%. This is because the speaker-dependent information is limited for unvoiced consonants, as discussed in Section IV-B. Thus, when there is no significant distortion between whisper and neutral non-unvoiced consonants, more emphasis would be placed on the MFCC-GMM system. Alternatively, when the whisper non-unvoiced consonant differs from the neutral version significantly, information from the unvoiced consonants can be helpful even if it is limited. Hence, in this case, more emphasis will be placed on unvoiced consonants. This result also confirms the potential improvement for whisper speaker recognition by combining scores from unvoiced and non-unvoiced consonant systems that are based on alternative mapping scale feature extraction methods.

Before determining an appropriate value of α, the threshold for confidence measurement is found by using a leave-one-out cross-validation procedure on all available read whisper data. Specifically, in each round, one speaker is chosen as the validation speaker, and the data from the remaining 27 speakers are used as development data. A threshold is found for each group of 27 speakers that can result in the lowest equal error rate for both high/low-performance whisper. The final threshold set is based on the average of these 28 thresholds.

Next, eight read whisper utterances from each speaker are selected as development data. Different sets of α are applied using the development data, and the α pair (0.4, 0.5) achieving the best performance for the dev-set data is applied for our final system in Fig. 12. When a whisper utterance is classified as low-performance whisper, α in (11) is set to 0.4, and when the utterance is classified as high-performance whisper, α is increased to 0.5. Finally, results from the MFCC and LFCC/EFCC-GMM systems are combined based on the prior α to obtain the overall final speaker ID result.

B. Experimental Results

Table VI summarizes the recognition results for closed-set speaker ID using neutral trained models and whispered test

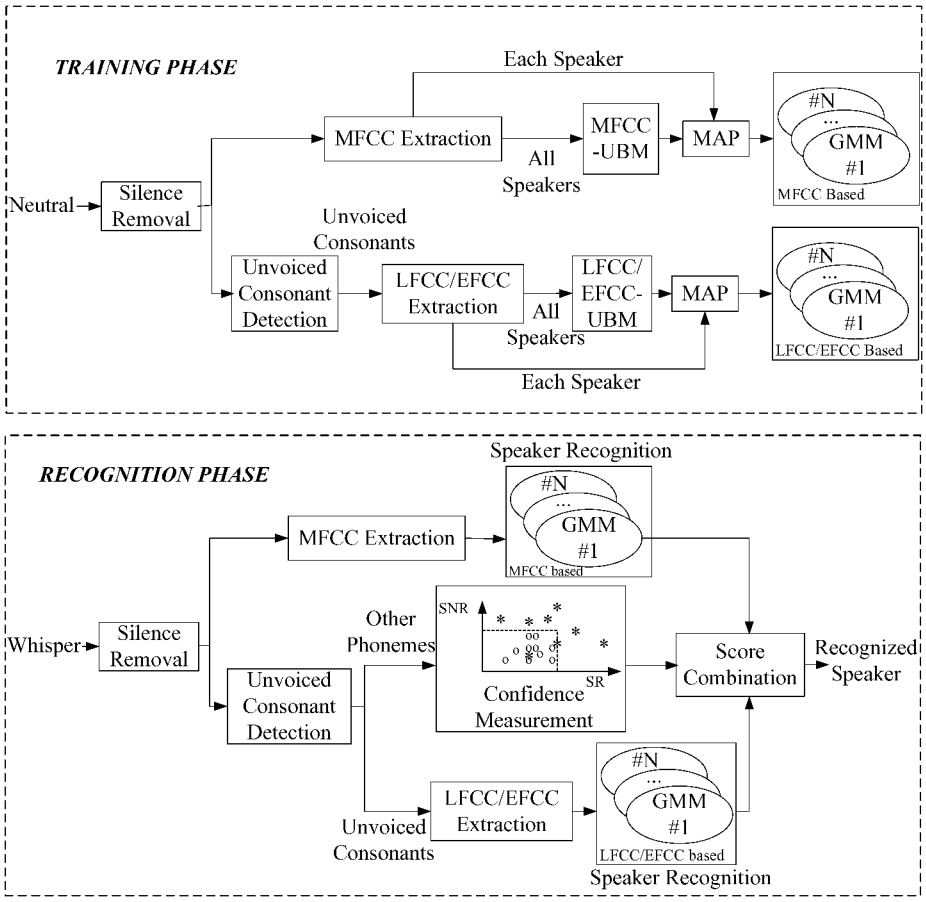


Fig. 12. System flow diagram for speaker recognition of whisper based on neutral trained GMMs.

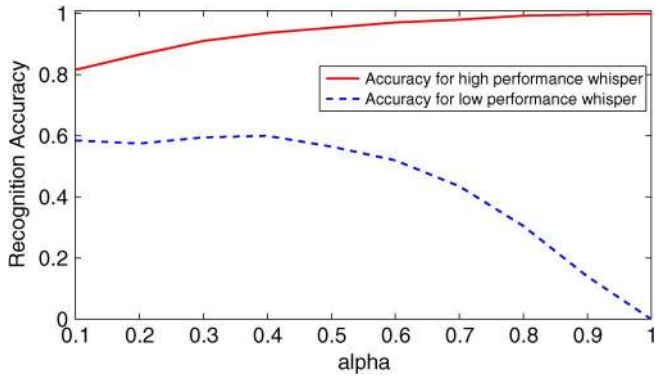


Fig. 13. Recognition accuracy of high/low-performance whisper of different alpha.

speech. Again, all speaker GMMs are trained with neutral speech, and test utterances are drawn from both the read and spontaneous whispered portion of our corpus, with a total of 961 read and 495 spontaneous test utterances. For the spontaneous part, the same hard threshold for the confidence space measurement and α remains the same as used in the read part. For each test utterance, the score from the MFCC-GMM and EFCC/LFCC-GMM systems are combined with the prior α that is determined from the results of the confidence space measurement. In order to demonstrate the effectiveness of confidence measurement, a fixed α of 0.5 for both MFCC and LFCC/EFCC (value determined experimentally) is applied to combine the score and the corresponding accuracies are listed in the last column in Table VI.

TABLE VI
RECOGNITION RESULTS FOR CLOSED-SET SPEAKER ID

Whisper	Feature vector	Accuracy	Accuracy (fixed *alpha*)
read	MFCC	79.29%	
read	MFCC+LFCC	88.35%	87.30%
read	MFCC+EFCC	88.14%	87.30%
spontaneous	MFCC	73.54%	
spontaneous	MFCC+LFCC	83.23%	82.83%
spontaneous	MFCC+EFCC	83.84%	83.23%

When MFCCs are used alone, speaker ID performance is 79.29% for read and 73.54% for spontaneous whispered speech. However, when the score of the MFCC-GMM system is combined with the score from either the EFCC or LFCC-GMM-based systems, a significant absolute improvement of 8.85%–10.30% is observed, which demonstrates the effectiveness of EFCC/LFCC feature processing for speaker ID of whispered speech. A relative improvement in error rate from 2.33%–8.27% is achieved when a different score weight for α is applied. These improvements demonstrate the efficiency of the confidence measurement. It is noted that further compensation processing can also be applied to whispered speech, according to the estimated whispered speech quality. A further investigation of new compensation strategies will be a focus of future work. The focus here, however, has been on combining results from alternative MFCC and LFCC/EFCC-GMM-based

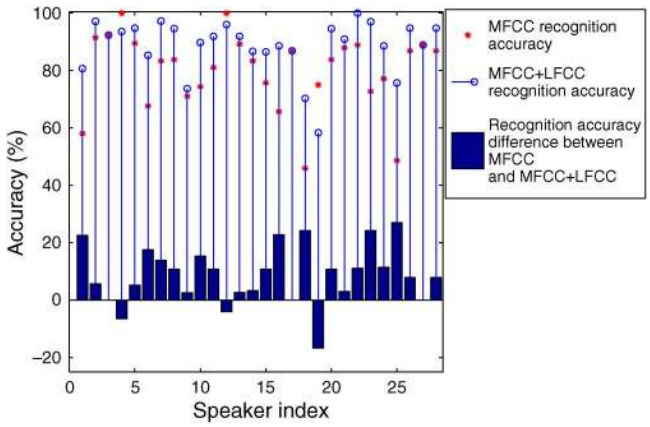


Fig. 14. Comparison of recognition accuracy of whisper between two systems for each speaker. Average improvement when the system helps: +11% for 25/28 speakers; average loss for 3/28 speakers when it hurts: −9%.

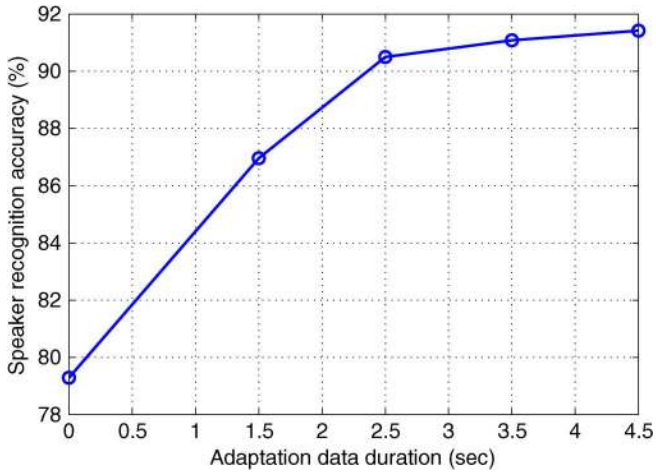


Fig. 15. Recognition accuracy of different duration of whisper adaptation data.

systems, as well as whisper quality selection for speaker ID. Since it is expected that the input audio stream will contain a mixture of both whispered and neutral speech, sustained speaker ID performance for neutral speech is also a goal. For each speaker, 40 neutral utterances are used for training and the remaining 477 utterances are used for testing. The proposed system achieves an accuracy of 99.16%, which confirms the system's effectiveness under the neutral/neutral train/test scenario. These results also confirm the seamlessness of the proposed speaker ID system for both neutral/whisper and neutral/neutral train/test scenarios.

To better demonstrate the consistency of the proposed system, the performance for each speaker using MFCC and MFCC+LFCC is also tabulated in Fig. 14. The performance differences, presented as black bars, are also shown (e.g., positive bars reflect improvement). When using MFCC+LFCC, improvement is achieved for both low- and high-performance whisper speakers. 22 of 28 speakers showed a measurable improvement; 3 speakers showed no improvement. Finally, 2 of 28 speakers showed slight decrease in performance, and one showed a large decrease in performance.

C. Comparison With Adaptation Method

This study has assumed that no whisper adaptation data is available, since most state-of-the-art speaker recognition systems are trained with only neutral speech. However, it would be useful to compare performance of the proposed system with a conventional MFCC-GMM system with adaptation data. In this study, the speaker dependent whisper model is obtained using MAP estimation using a neutral trained 64-mixture UBM. The duration of the amount of whisper adaptation data was varied from 0 to 4.5 s. Again, closed-set speaker recognition is conduced on read whisper speech from all 28 speakers. The results are shown in Fig. 15.

From Fig. 15, it can be seen that adaptation whisper data is helpful for improving system performance, especially when 2.5 s or more adaptation speech is available. For example, when the adaptation whisper data duration is 1.5 s, the recognition accuracy is 86.96%, 1.39% below the accuracy of the proposed MFCC+LFCC system. With increasing amounts of adaptation data, the accuracy reaches 91.42% when 4.5 s of whisper adaptation data is available for each speaker. This result shows that, depending on the duration of the whisper adaptation data, the MFCC+LFCC/EFCC system can still outperform a conventional MAP adapted MFCC-GMM systems. We also compare the performance using the adapted models for the three speakers in Section VI-B, whose performance has been degraded using the proposed system. The average accuracy for the three speakers using the 4.5 s adapted model is 98.70%, which is significantly improved compared to 80.43% using the proposed system. For the remaining 25 speakers, the average accuracy for speaker ID is 90.65% using the 4.5 s whisper adapted models, compared to 89.18% using the proposed system. These results show that the proposed system can provide comparable results for a majority of speakers to the whisper adapted models.

VII. Summary and Conclusion

Whisper is an alternative speech production mode that is employed by individuals for communication in public circumstances to protect personal privacy. However, the performance of traditional MFCC-GMM speaker ID systems degrade rapidly due to the significant differences in speech production between whispered and neutral speech. The goal of this study therefore has been to develop a seamless closed-set speaker recognition system that works for both neutral/whisper mismatched and neutral/neutral matched scenarios.

An MFCC-GMM baseline was first formulated, and it was determined that performance of traditional MFCC-GMM speaker ID systems varies significantly among speakers using whispered speech. In order to determine the cause for this variation, the KL divergence between GMMs trained on unvoiced consonant and on other phonemes was compared for each speaker. The result confirmed that differences between whisper and neutral speech are concentrated in phonemes other than unvoiced consonants. Based on the spectral properties of unvoiced consonants, feature vectors based on linear and exponential mapping functions were applied, resulting in an improvement in recognition accuracy. Next, a set of acoustic analysis steps were conducted on whispered non-unvoiced consonants (e.g., whisper version of vowels, liquids, glides, diphthongs, and nasals). A 2-D confidence space was introduced based on the

three metrics: SNR, spectral tilt, and $ratio^{f}_{1-2vs1-8}$, to assess the quality of whisper for speaker ID. Experimental evaluations showed that, despite imperfect classification of whispered speech as either high- or low-performance whisper, useful portion is classified correctly with high confidence. When combining MFCC+LFCC/EFCC-GMM and the confidence space measurement, the proposed system achieved an absolute improvement of 8.85%–10.30% in speaker recognition, compared to the MFCC-GMM baseline. From a total 961 test read whisper utterances, the best closed-set speaker ID performance obtained is 88.35%, and for the total 495 spontaneous whisper utterances, 83.84% accuracy is achieved. This result was also compared to the accuracy obtained using a neutral baseline model adapted with whispered speech, and results showed that when the whisper adaptation data is set for 1.5 s for each speaker, the proposed system outperforms the conventional MAP-adapted model.

This study has therefore established a viable approach to improve speaker recognition when test evaluations require whispered speech, and only neutral speech is available for training. It is important to note that whispered speech is more likely to be encountered as embedded utterance segments with neutral speech in the surrounding time domain. As such, it is important to develop speaker ID systems that sustain performance as the speaker alters their speaker mode. In the next phase, future work could consider compensation for low performance non-unvoiced consonants within whisper.

REFERENCES

[1] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: Whisper through shouted," in *Proc. Interspeech*, 2007, pp. 2289–2292.

[2] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Commun.*, vol. 45, no. 2, pp. 139–152, 2005.

[3] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Med. Eng. Phys.*, vol. 24, no. 7–8, pp. 515–520, 2002.

[4] S. T. Jovicic, "Formant feature differences between whispered and voiced sustained vowels," *Acustica-Acta*, vol. 84, no. 4, pp. 739–743, 1998.

[5] M. Matsuda and H. Kasuya, "Acoustic nature of the whisper," in *Proc. Eurospeech*, 1999, pp. 133–136.

[6] Q. Jin, S. S. Jou, and T. Schultz, "Whispering Speaker Identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2007, pp. 1027–1030.

[7] X. Fan and J. H. L. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition," in *Proc. ISCA Interspeech*, 2008, pp. 1313–1316.

[8] X. Fan and J. H. L. Hansen, "Speaker Identification with whispered speech based on modified LFCC parameters and feature mapping," in *Proc. IEEE ICASSP*, 2009, pp. 4553–4556.

[9] X. Fan and J. H. L. Hansen, "Speaker Identification for whispered speech using modified temporal patterns and MFCCs," in *Proc. ISCA Interspeech*, 2009, pp. 896–899.

[10] T. Hueber, G. Chollet, B. Denby, and M. Stone, "Acquisition of ultrasound, video and acoustic speech data for silent-speech interface application," in *Int. Seminar Speech Production*, Strasbourg, France, 2008, pp. 365–369.

[11] T. Hueber, E. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface," in *Proc. ISCA Interspeech*, 2009.

[12] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-Audible Murmur (NAM) recognition," *IEICE Trans. Inf. Syst.*, vol. E89-D, pp. 1–4, 2006.

[13] S. Jovicic and Z. Saric, "Acoustic analysis of consonants in whispered speech," *J. Voice*, vol. 22, pp. 263–274, 2008.

[14] L. Gavidia-Ceballos, "Analysis and modeling of speech for laryngeal pathology assessment," Ph.D. dissertation, RSPL: Robust Speech Process. Lab., Duke Univ., Durham, NC, 1995.

[15] J. H. L. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 300–313, Mar. 1998.

[16] L. Gavidia-Ceballos and J. H. L. Hansen, "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 4, pp. 373–383, Apr. 1996.

[17] W. Meyer-Eppler, "Realization of prosodic features in whispered speech," *J. Acoust. Soc. Amer.*, vol. 29, pp. 104–106, 1957.

[18] I. Thomas, "Perceived pitch of whispered vowels," *J. Acoust. Sco. Amer.*, vol. 46, pp. 468–470, 1969.

[19] I. Eklund and H. Traunmuller, "Comparative study of male and female whispered and phonaed versions of the long vowels of Swedish," *Phonetica*, vol. 54, pp. 1–21, 1996.

[20] K. J. Kallail and F. W. Emanuel, "Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects," *J. Speech Hear. Res.*, vol. 27, pp. 245–251, 1984.

[21] C. Zhang and J. H. L. Hansen, "Advancement in whisper-island detection with normally phonated audio streams," in *Proc. ISCA Interspeech*, 2009, pp. 860–863.

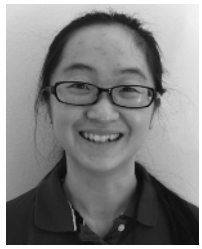
[22] H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1783–1752, 1990.

[23] R. J. Mammone, X. Zhang, and R. R. Ramachandran, "Robust speaker recognition: A feature based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 58–71, Sep. 1996.

[24] S. Kullback, *Information Theory and Statistics*. Mineola, NY: Dover, 1968.

[25] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans Speech Audio Process.*, vol. 8, no. 4, pp. 429–442, Jul. 2000.

[26] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, Jul. 1988.



Xing Fan received the B.S. degree in acoustics from Nanjing University, Nanjing, China in 2005. She is currently pursuing the Ph.D. degree in electrical engineering at the University of Texas at Dallas, Richardson.

She has been working as a Research Assistant with the Center for Robust Speech Systems (CRSS), University of Texas at Dallas. Her research interests include speech signal processing, machine learning, and pattern recognition.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982. and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, GA, in 1983 and 1988, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado at Boulder (1998–2005), where he co-founded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 51 (22 Ph.D., 29 M.S./M.A.) thesis candidates.

He is author/coauthor of 380 journal and conference papers and eight textbooks in the field of speech processing and language technology, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), co-editor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was named IEEE Fellow in 2007 for contributions in "Robust Speech Recognition in Stress and Noise," and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee (2005–2008; 2010–2013; elected Chair-elect in 2010), and Educational Technical Committee (2005–2008; 2008–2010). Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/2006), Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the IEEE Signal Processing Magazine (2001–2003). He has also served as a Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the ISCA (International Speech Communications Association) Advisory Council. In 2010, he was recognized as an ISCA Fellow, for contributions on "research for speech processing of signals under adverse conditions." He was recipient of The 2005 University of Colorado Teacher Recognition Award as voted on by the student body. He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and served as Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX.