CASA-Based Robust Speaker Identification

Xiaojia Zhao, *Student Member, IEEE*, Yang Shao, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Conventional speaker recognition systems perform poorly under noisy conditions. Inspired by auditory perception, computational auditory scene analysis (CASA) typically segregates speech by producing a binary time–frequency mask. We investigate CASA for robust speaker identification. We first introduce a novel speaker feature, gammatone frequency cepstral coefficient (GFCC), based on an auditory periphery model, and show that this feature captures speaker characteristics and performs substantially better than conventional speaker features under noisy conditions. To deal with noisy speech, we apply CASA separation and then either reconstruct or marginalize corrupted components indicated by a CASA mask. We find that both reconstruction and marginalization are effective. We further combine the two methods into a single system based on their complementary advantages, and this system achieves significant performance improvements over related systems under a wide range of signal-to-noise ratios.**

***Index Terms*—Computational auditory scene analysis (CASA), gammatone frequency cepstral coefficient (GFCC), ideal binary mask, robust speaker identification.**

I. Introduction

A SPEAKER recognition system, performing either speaker identification (SID) or speaker verification (SV) tasks, comprises three processes: feature extraction, speaker modeling, and decision making using pattern classification methods [3], [8]. Typically, extracted speaker features are short-time cepstral coefficients such as Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP) coefficients, or long-term features such as prosody [32]. For speaker modeling, Gaussian mixture models (GMMs) are widely used to describe feature distributions of individual speakers [26]. Recognition decisions are usually made based on likelihoods of observing feature frames given a speaker model. Such systems usually do not perform well under noisy conditions [10], [31] because extracted features are distorted by noise, causing mismatched likelihood calculation.

To tackle this robustness problem, speech enhancement methods that are widely used in speech recognition, such as spectral subtraction, have been explored for robust speaker recognition [23], [38]. However, these methods do not perform well when noise is nonstationary. RASTA filtering [11] and cepstral mean normalization (CMN) [9] have been used in speaker recognition but they are mainly intended for convolutive noise. Studies of robust speech recognition on Aurora [20] have yielded an advanced front-end feature extraction algorithm (AFE) [35], which is standardized by the European Telecommunication Standards Institute (ETSI). ETSI-AFE derives robust MFCC features using a set of sophisticated front-end processes, including speech activity detection and Wiener filtering. An alternative approach to feature enhancement seeks to improve robustness by modeling noise and combining it with clean speaker models [17], [39].

On the other hand, similar to speech recognition tasks, human listeners perform robustly in speaker recognition tasks [28]. The human ability to function well in noisy acoustic environments is due to a perceptual process termed auditory scene analysis (ASA) [2]. Inspired by ASA research, computational auditory scene analysis (CASA) aims to organize sound based on ASA principles [37]. The robust performance of the auditory system motivates us to explore CASA for robust speaker recognition.

In this paper, we propose a robust speaker identification system by using CASA as a front-end to perform speech segregation. The output of CASA segregation is in the form of a binary time–frequency (T-F) mask that indicates whether a particular T-F unit is dominated by speech or background noise. We first propose new speaker features, gammatone feature (GF) and gammatone frequency cepstral coefficients (GFCC), based on an auditory periphery model. Specifically, a GF is first obtained from a bank of gammatone filters. Then, GFCC is derived from GF by a cepstral analysis. We show that GFCC achieves an SID level of performance in noisy environments that is significantly better than MFCC. The proposed system has two modules. To account for the deviations of noisy features from clean ones, the first module enhances the GF by reconstructing corrupted components indicated by a CASA-generated binary T-F mask. The second module performs bounded marginalization on the noisy GF. Each module yields substantial improvement over baseline SID systems. As the two modules perform well in different conditions, we propose a combined system integrating these two modules.

The rest of the paper is organized as follows. Section II describes the overall system architecture. Auditory feature extraction and binary mask estimation are discussed in Section III. Sections IV and V introduce the reconstruction module and the marginalization module, respectively. The two modules are combined in Section VI. SID evaluations and comparisons are presented in Section VII. Further discussions are given in Section VIII.

Manuscript received September 23, 2010; revised March 04, 2011; accepted January 17, 2012. Date of publication February 03, 2012; date of current version March 21, 2012. This work was supported in part by the Air Force Research Laboratory (AFRL) as a subcontractor to RADC, Inc. under Grant FA8750-09-C-0067, as well as an AFRL Grant (FA8750-04-1-0093) and in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-1-0155. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nestor Becerra Yoma.

X. Zhao and Y. Shao are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: zhaox@cse.ohio-state.edu; shao.19@osu.edu).

D. Wang is with the Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASL.2012.2186803

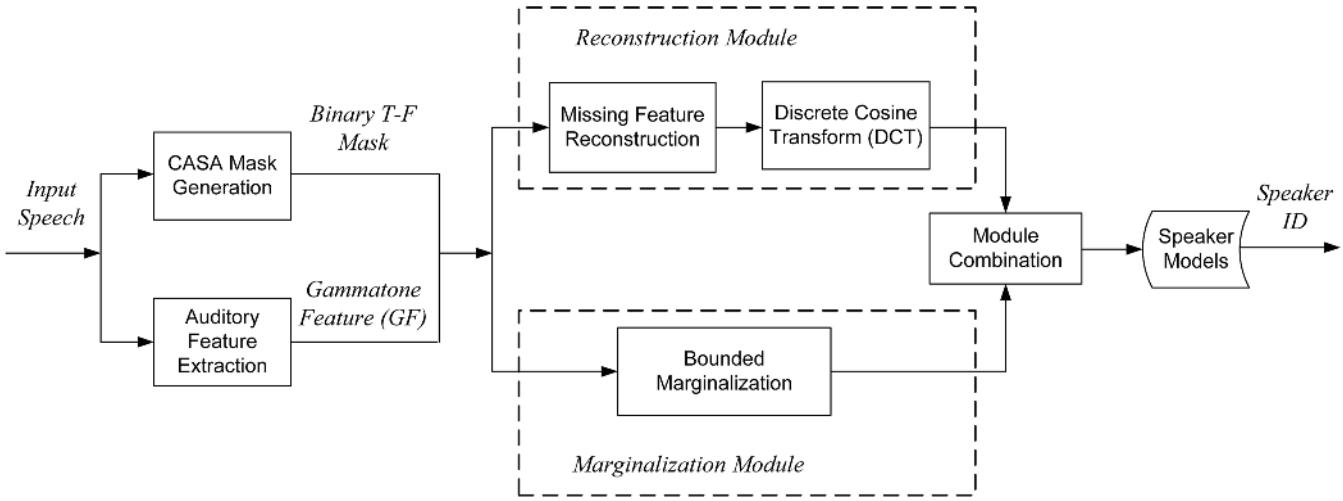

Fig. 1. Schematic diagram of a CASA-based robust speaker identification system.

II. System Overview

The proposed system uses CASA as a front-end processor for robust SID. Fig. 1 presents the diagram of the overall system. Input speech is decomposed using a gammatone filterbank and subsequent time windowing to generate a time sequence of GFs. This T-F analysis results in a cochleagram [37], which is a two-dimensional representation of the input signal. Simultaneously, we feed the input signal to a CASA system that computes a binary mask corresponding to the target speech [13]. Elements of this mask correspond to T-F units in the cochleagram, with 1 indicating that the corresponding T-F unit is dominated by target and 0 by noise. The binary mask and GFs are fed to both the reconstruction module and the marginalization module.

In the reconstruction module, the noise-corrupted components indicated by the CASA mask are reconstructed using a speech prior [24] and the enhanced GF is converted to the cepstral domain by discrete cosine transform (DCT). Subsequently, the obtained cepstral feature, GFCC, is used in conjunction with trained speaker models to derive the underlying speaker identity. In the marginalization module, there is no need for missing feature reconstruction. Bounded marginalization is performed on the noisy GF directly with the CASA mask providing the information of which T-F units are corrupted and hence marginalized.

Each module provides an SID system by itself. Our experiments suggest that the reconstruction module and the marginalization module work well in different conditions. To leverage their respective advantages, our combined system assigns the input signal to both modules and integrates the individual outputs to make the final decision. Note that the two modules as well as the combined system operate on a per utterance basis.

III. Feature Extraction and Mask Estimation

In this section, we describe how to extract GF and GFCC features from the cochleagram, and compute a CASA mask.

A. Auditory Features

Our system first performs auditory filtering by decomposing an input signal into the T-F domain using a bank of gammatone filters. Gammatone filters are derived from psychophysical and physiological observations of the auditory periphery and this filterbank is a standard model of cochlear filtering [21]. We use a bank of 64 filters whose center frequencies range from 50 Hz to 4000 Hz or 8000 Hz depending on the sampling frequency of speech data. Since the filter output retains the original sampling frequency, we decimate fully rectified 64-channel filter responses to 100 Hz along the time dimension. This yields a corresponding frame rate of 10 ms, which is used in many short-time speech feature extraction methods. The magnitudes of the decimated outputs are then loudness-compressed by a cubic root operation

$$G_m[i] = \left||g|_{decimate}\,[i,m]\right|^{1/3}, \quad i = 0 \ldots N-1, \quad m = 0 \ldots M-1. \tag{1}$$

Here, $N = 64$ refers to the number of frequency (filter) channels. M is the number of time frames obtained after decimation. The resulting responses $G_m[i]$ form a matrix, representing the T-F decomposition of the input. This T-F representation is a variant of cochleagram. Note that, unlike the linear frequency resolution of a spectrogram, a cochleagram provides a finer frequency resolution at low frequencies than at high frequencies. Fig. 2 shows a cochleagram and a spectrogram of an utterance. Darker regions represent stronger energy. Note the difference in energy-concentrated regions below 1000 Hz between these two T-F representations. We base our subsequent processing on the cochleagram representation.

We call a time slice of the above matrix gammatone feature (GF), and use $G[i]$ to denote its ith channel. Time index m is dropped for simplicity. Here, a GF vector comprises 64 frequency components. Note that the dimension of a GF vector is larger than that of MFCC vectors used in a typical speaker recognition system. Additionally, because of the frequency overlap among neighboring filter channels, GF components are correlated with each other. In order to reduce dimensionality and de-correlate the components, we apply a DCT to a GF. We call the resulting coefficients gammatone frequency cepstral coefficients (GFCCs) [29], [30]. Specifically, cepstral

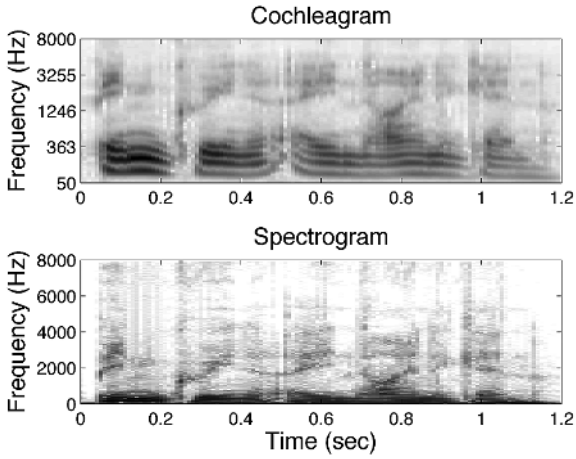

Fig. 2. Illustrations of a cochleagram (top) and a spectrogram (bottom) of a clean speech utterance. Note the asymmetric frequency resolution at low and high frequencies in the cochleagram.

coefficients, $C[j]$, $j = 0 \ldots N-1$, are obtained from a GF as follows:

$$C[j] = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} G[i] \cos\left(\frac{j\pi}{2N}(2i+1)\right), \quad j = 0 \ldots N-1. \tag{2}$$

Note that the zeroth-order coefficient summates all the GF components. Thus, it relates to the energy of a GF vector.

Rigorously speaking, the newly derived coefficients are not cepstral coefficients because a cepstral analysis requires a log operation between the first and the second frequency analysis for the deconvolution purpose [19]. Here, we call them cepstral coefficients because of the functional similarities between the above transformation and that of a typical cepstral analysis in the derivation of MFCC.

B. CASA-Based Mask Estimation

As described earlier, a cochleagram is a T-F representation of a signal. With such a representation, a binary T-F mask furnishes the crucial information about whether a T-F unit is dominated by target speech or background noise. As a main computational goal of CASA, an ideal binary mask (IBM) is a binary matrix defined as follows [36]:

$$IBM(t,f) = \begin{cases} 1, & \text{if } SNR(t,f) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

$IBM(t,f)$ is indexed by time t and frequency f. $SNR(t,f)$ refers to the local signal-to-noise ratio (SNR) (in dB) for the T-F unit in time frame t and frequency channel f. Given premixed target and interference signals, the IBM can be readily constructed. The IBM concept is motivated by the auditory masking phenomenon [18], and is the optimal binary mask in terms of SNR gain [16].

To estimate the IBM from an input mixture, we employ a recent CASA system that performs feature-based classification [13]. First, we estimate the pitch of the speech signal at each frame using a multipitch tracking algorithm [12]. This algorithm formulates multipitch tracking as a hidden Markov model

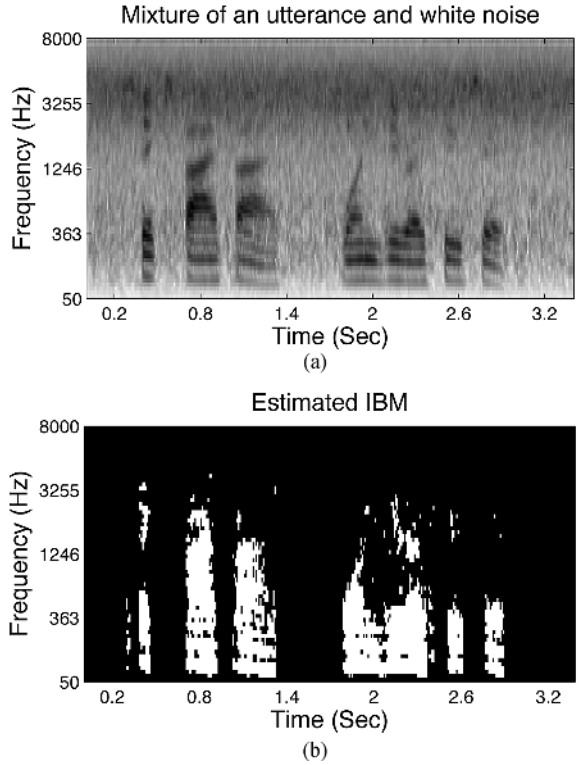

Fig. 3. Illustration of estimated IBM produced by a voiced speech segregation system. The top plot shows the cochleagram of an utterance mixed with white noise at 0-dB SNR. The bottom plot presents an estimated IBM from the mixture in the top plot, where 1 is shown as white and 0 as black.

(HMM), which can produce up to two pitch points at each frame. As we deal with noises that are mostly aperiodic, the multipitch tracker tends to output at most one pitch per frame. Given an estimated pitch, a six-dimensional pitch-based feature vector is extracted for each T-F unit [13]. These features are fed to an MLP (multilayer perceptron) classifier, whose output can be interpreted as the posterior probability of a T-F unit being target dominant. The desired output during MLP training is the IBM. Note that we take the binarized MLP output as the resulting CASA mask without using a subsequent segmentation and grouping stage in the original system of [13].

Fig. 3 shows an estimated IBM for a noisy speech utterance. If input SNR is given, an SNR-dependent MLP can be trained to estimate the IBM. Otherwise, one can train multiple MLPs at different SNRs, and select the MLP whose corresponding SNR is closest to the estimated input SNR. In this paper, the latter is adopted as we assume no prior knowledge of input SNR. More details about MLP training will be provided in Section VII-A.

IV. Reconstruction Module

In speaker recognition, the probability distribution of an extracted feature vector X produced by a speaker λ is typically modeled as a GMM [26], parameterized by diagonal covariance matrices. Under noisy conditions, the aforementioned speech segregation system produces a binary T-F mask that indicates whether a GF component is speech dominant or noise dominant. The former one is regarded as reliable since the system has more speech information in the speaker models while the latter one is

deemed missing. Thus, the feature vector is partitioned into reliable components X_r, and unreliable ones X_u:

$$X = \begin{bmatrix} X_r \\ X_u \end{bmatrix}. \tag{4}$$

In order to enhance a noise corrupted GF, we first reconstruct its missing components from a speech prior model, which is similar to a universal background model in speaker verification [25]. Specifically, the speech prior $p(X)$ is modeled as a large GMM [24], and obtained from pooled training data:

$$p(X) = \sum_{k=1}^{K} p(k)p(X \mid k) \tag{5}$$

where K is the number of mixture components, and k denotes the index and $p(k)$ the prior probability of a mixture component. $p(X|k)$ is the kth Gaussian distribution with a mean vector μ_k and a diagonal covariance σ_k^2. Given a binary mask, the components of the mean and variance of each Gaussian are also split into reliable and unreliable ones. We then calculate the *a posteriori* probability of the kth component given reliable GF components as

$$p(k \mid X_r) = \frac{p(k)p(X_r \mid k)}{\sum_{k=1}^{K} p(k)p(X_r \mid k)}. \tag{6}$$

As shown in [4], [34], the unreliable components are estimated as the expected value or the mean conditioned on X_r:

$$\hat{X}_u = \sum_{k=1}^{K} p(k \mid X_r)\mu_{u,k} \tag{7}$$

where $\mu_{u,k}$ refers to the mean vector of the unreliable components of the kth Gaussian in the speech prior. The reliable components are retained in the reconstruction. Since GF is an energy-based feature, the underlying target signal is expected to be smaller than the mixture value. Therefore, we replace a reconstructed value with the observed value if the former is larger.

As shown in the above equations, the quality of reconstruction is largely determined by the amount of reliable speech information. With little reliable information, the quality of recognition is expected to be very poor. Therefore, we introduce a frame selection step in the reconstruction module to choose relatively clean frames, when there are plenty frames available for recognition. Some criterion such as frame level SNR or the number of reliable units is needed for selection, and details will be provided in Section VII-D.

With the reconstructed GF, we convert it into GFCC by applying DCT. GFCC is a speaker feature that can be directly used for recognition in conjunction of trained speaker models as described in Section III-A.

V. Marginalization Module

An alternative approach to reconstruction is marginalization, which has shown good performance in robust speech recognition [4] and has been applied to robust speaker recognition [7], [31]. The main idea behind marginalization is to base recognition decisions on reliable components; in other words, we want to marginalize unreliable components. With GMM speaker models and diagonal covariance matrices, we have

$$\begin{aligned} p(X_r \mid \lambda) &= \int_{-\infty}^{\infty} p(X_r, X_u \mid \lambda) dX_u \\ &= \int_{-\infty}^{\infty} \sum_{k=1}^{K} p(k)p(X_r, X_u \mid k) dX_u \\ &= \sum_{k=1}^{K} p(k)p(X_r \mid k) \int_{-\infty}^{\infty} p(X_u \mid k) dX_u \\ &= \sum_{k=1}^{K} p(k)p(X_r \mid k). \end{aligned} \tag{8}$$

In the above equation, an unreliable feature dimension integrates to 1 and the likelihood calculation reduces to a simple case where the feature dimensions of reliable T-F units are inserted into each speaker model to get the likelihood of a frame.

Although from the unreliable T-F units we cannot precisely predict the underlying target feature value, the feature value should be within the range from 0 to the observed value as a GF feature is derived from the cubic root operation [see (1)]. This analysis provides a more accurate range of integration than that from minus infinity to positive infinity in (8). Utilizing the tighter range leads to bounded marginalization [4], described as follows where "low" and "high" define the range:

$$\begin{aligned} p(X_r \mid \lambda) &= \int_{low}^{high} p(X_r, X_u \mid \lambda) dX_u \\ &= \int_{low}^{high} \sum_{k=1}^{K} p(k)p(X_r, X_u \mid k) dX_u \\ &= \sum_{k=1}^{K} p(k)p(X_r \mid k) \int_{low}^{high} p(X_u \mid k) dX_u. \end{aligned} \tag{9}$$

Consistent with earlier studies [4], [7], we have found that bounded marginalization produces substantially better recognition performance than full marginalization. Therefore, we employ bounded marginalization on GF features. It is worth emphasizing that this marginalization method operates in the spectral domain, whereas the reconstruction method described in Section IV performs recognition in the cepstral domain.

VI. Combined System

Between the reconstruction module and the marginalization module, we expect the former to perform better at high SNRs as it is well known that cepstral features outperform spectral features in recognition [6], [33]. On the other hand, marginalization is expected to perform better in low SNR conditions, as reconstruction based on few reliable T-F units likely has poor quality. Also, bounded marginalization makes use of some information from unreliable T-F units. These differing performance trends are indeed confirmed by the evaluation results presented in the next section. To utilize the relative advantages, we combine them into one system.

In our study, we have noticed that when a module makes a recognition mistake, the underlying target speaker tends to have

a top ranked score although it is not the highest. Meanwhile, wrong identities from these two modules tend not to agree. Motivated by this observation, we simply use a linear combination.

We derive an SID score vector for each frame by feeding the frame signal to each speaker model. Note that each element of this vector is a log-likelihood corresponding to a particular speaker model. An utterance level score vector is derived by adding frame level log-likelihood score vectors. After integrating SID scores from all the available frames, each module outputs a score vector with the number of elements equal to the total number of the speaker models. As the scores from the two modules may not be on the same scale, normalization should be applied before adding them together. We perform the following simple normalization:

$$\hat{s}_{Module}(\lambda) = \frac{s_{Module}(\lambda) - \min_{\lambda}\left(s_{Module}(\lambda)\right)}{\max_{\lambda}\left(s_{Module}(\lambda)\right) - \min_{\lambda}\left(s_{Module}(\lambda)\right)} \quad (10)$$

where $s_{Module}(\lambda)$ and $\hat{s}_{Module}(\lambda)$ denote the original and normalized score vectors of an individual module respectively. The SID score of the combined system is given as follows:

$$s(\lambda) = \hat{s}_{REC}(\lambda) + \hat{s}_{MAR}(\lambda). \quad (11)$$

We refer to the frames containing at least one reliable T-F unit as "active frames." The frames containing no reliable unit are either unvoiced speech mixed with noise or voiced speech completely masked by noise. Our study shows that unvoiced speech plays a relatively minor role in speaker recognition and our CASA-mask estimation algorithm cannot separate unvoiced speech. Completely masked voiced speech provides little information for SID and it seems reasonable to ignore these frames. Therefore, we only feed active frames to the two modules.

VII. Evaluation and Comparison

In this section, we systematically evaluate the noise robustness of the proposed SID methods. We also compare the performance of our system with baseline systems using the conventional MFCC feature and the ETSI-AFE feature. In addition, we compare with a related robust SID system by Pullella *et al.* [23].

A. Experiment Setup

We employ speech material (one-speaker detection, cellular data) from the 2002 NIST Speaker Recognition Evaluation corpus [22], which is a standard dataset for automatic speaker recognition (particularly verification). The speaker dataset contains 330 speakers. Each speaker has a roughly 2-minute-long telephone recording sampled at 8 kHz for training. It is divided into 5-s-long pieces, and 2 of them are included in the test set, 2 in the development set and the remaining ones in the training set. To study how the proposed system performs under different types of noisy conditions, the test utterances are mixed with multitalker babble noise which is nonstationary, speech shape noise (stationary), and factory noise (nonstationary). Each noise is mixed with telephone speech at various SNR levels from -6 dB to 18 dB at 6-dB intervals. Note that the test utterances are different from the training ones.

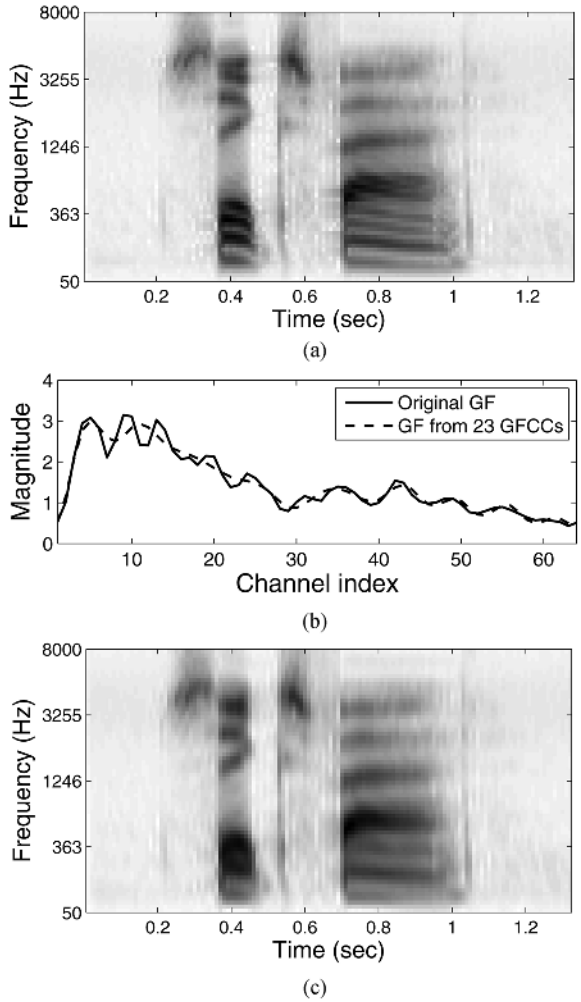

Fig. 4. Illustrations of energy compaction by GFCCs. Plot (a) shows a cochleagram of an utterance. Plot (b) shows a GF frame at 1 s of (a). The original GF is plotted as the solid line and the resynthesized GF by 23 GFCCs is plotted as the dashed line. Plot (c) presents the resynthesized cochleagram from (a) using 23 GFCCs.

Sixty-four dimensional GF is extracted to model speaker dependent characteristics. To reconstruct the noisy GF, a speech prior with 2048 Gaussian components is trained using all the pooled training data. The reconstructed GF is converted to GFCC using DCT. Each speaker model is adapted from a 1024-component universal background model (UBM) trained using all the training data [27]. Compared with individually trained GMMs, this GMM-UBM approach scores much faster and is more discriminative.

As the NIST dataset contains telephone speech, little speech information exists below 200 Hz. Therefore, we only use features above 200 Hz. In the gammatone filterbank, the ten lowest channels correspond to frequencies below 200 Hz and thus GF consists of channels 11–64 (i.e., 54 channels). As confirmed using the development set, excluding the low-frequency channels increases SID performance.

For MLP training, we randomly select 50 utterances from the training set and mix them with speech shape, factory, babble, and white noises at SNR levels from -12 dB to 18 dB with 6-dB increments. At each SNR, an SNR-specific MLP is trained. In addition, a generic MLP is trained by pooling mixtures from all

SNR levels. Given a test speech signal, the generic MLP is used to generate a binary mask, from which we estimate the input SNR (during voiced intervals). For separation, we choose the MLP whose training SNR is closest to the estimated SNR.

B. GFCC Dimensions and Dynamic Features

In the reconstruction module, when converting 64-dimensional GF to GFCC, keeping all the 64 dimensions of GFCC may not be necessary. After inverting DCT of GFCC, we find that the lower 23-order coefficients capture almost all the GF information and the coefficients above the 23th have values close to 0, which means that they provide negligible information (see also [30]). As an illustration, Fig. 4(a) shows the cochleagram of an utterance, Fig. 4(b) shows a comparison of a GF frame at 1 s of Fig. 4(a) and the resynthesized GF from the first 23 GFCC coefficients, and Fig. 4(c) presents the resynthesized cochleagram from the top plot using only the 23 coefficients. As can be seen from the figure, the lower 23-order GFCCs largely retain the information in 64-dimensional GFs. This is due to the "energy compaction" property of DCT [19]. Additionally, the zeroth cepstral coefficient corresponds to the energy of the whole frame, which is susceptible to noise corruption. Our experiments using the IBM for separation show that removing the zeroth coefficient improves the SID performance significantly. Hence, in the later experiments we will use 22-dimensional GFCCs.

Since a typical speaker recognition system uses MFCCs and their first-order (delta) dynamic coefficients, it is reasonable to study how GFCC dynamic features fare for recognition. GFCCs with 22 dimensions have shown good SID performance in our experiments. After appending 22-dimensional dynamic features, we find that the performance improvement is not significant. Therefore, we use 22-dimensional static GFCCs as speaker features in the reconstruction module.

C. Baseline Comparisons

To show the utility of GFCC as speaker features, we choose MFCC and ETSI-AFE as baseline features. ETSI-AFE is essentially enhanced MFCC features. Our experiments suggest that MFCC without delta or acceleration features performs better. This is probably because without noise reduction, the delta and acceleration features are very noisy and cannot encode the underlying dynamic speaker information. However, ETSI-AFE with delta features works better than static features. Therefore, we choose static MFCC features and ETSI-AFE with delta features as two baselines. For the GFCC baseline, we directly derive noisy GFCC features out of a mixture without separation or reconstruction. In this way, we could directly evaluate the effectiveness of GFCC as a new speaker feature. To make a fair comparison, since the GFCC feature has 22 dimensions, we also derive 22-dimensional MFCCs in addition to the commonly used 12-dimensional version (after removing the zeroth coefficient). As mentioned in Section VII-A, we only use GF features above 200 Hz. This is also the case for MFCC features. As for ETSI-AFE features, we use the default frequency range as it is unclear how to adjust the frequency range.

Fig. 5 gives the SID accuracies of different baseline systems with respect to SNR. When performing SID, we only consider

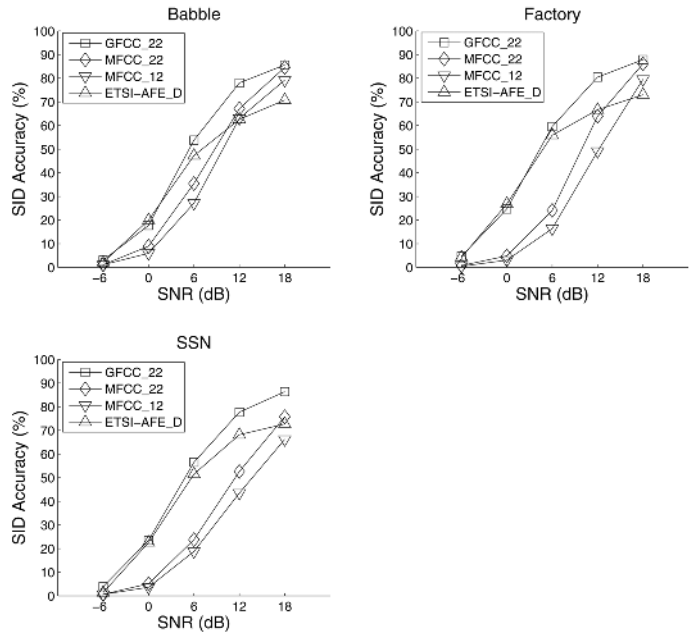

Fig. 5. SID performance of different baseline systems for three noises.

active frames. The results in the figure show that the GFCC baseline on average gives significantly better performance than the other baselines for all three noises. This indicates that the GFCC feature has more robustness in noisy conditions. ETSI-AFE_D (_D indicates delta features) works better than 12-dimensional MFCC features. After we increase MFCC dimensions to 22, the same dimensionality as GFCC, MFCC features yield closer performance to ETSI-AFE_D but still underperform GFCC features.

D. Evaluation Results

Now we present SID results of the proposed methods using estimated IBM. We also compare the performance of the individual modules and the combined system. In the frame selection step of the reconstruction module, we use as the selection criterion the smaller of half of the frequency channels (i.e., 27 for the NIST dataset—see Section VII-A) and the median number of reliable T-F units of all active frames for a noisy speech utterance. Given an active frame, it will be selected if its number of reliable units is greater than the criterion.

Table I shows the SID performances of the three methods: the reconstruction module, the marginalization module, and the combined system. As shown in the table, at high SNR conditions, particularly at 18 dB, the reconstruction module with GFCC performs well, better than GF plus bounded marginalization that operates in the spectral domain. On the other hand, the marginalization module performs consistently better under low SNR conditions. This suggests that, when there are relatively many reliable T-F units, reconstructing unreliable ones and using GFCC features yield performance advantages. On the contrary, if there are few reliable T-F units, bounded marginalization in the spectral domain is a more effective strategy. We should point out that, in terms of computational complexity, the reconstruction module is faster as it uses 22-dimensional GFCC features, as opposed to 54-dimensional GF features used in the marginalization module. Also, the integration operation

TABLE I
SID ACCURACY (%) OF THE PROPOSED METHODS. REC DENOTES THE RECONSTRUCTION MODULE, MAR THE MARGINALIZATION MODULE, AND CMB THE COMBINED SYSTEM

Babble	-6dB	0dB	6dB	12dB	18dB	Avg.
REC	13.94	47.58	70.45	79.85	86.21	59.61
MAR	21.21	54.85	72.12	80.15	83.79	62.42
CMB	31.97	70.30	82.12	87.88	90.61	72.58
Factory	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
REC	19.55	46.21	68.94	78.48	86.52	59.94
MAR	28.64	53.18	65	69.24	81.21	59.45
CMB	38.33	67.73	77.88	83.64	89.09	71.33
SSN	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
REC	15.76	40.76	66.21	78.64	85.76	57.43
MAR	25.91	55	72.73	77.73	82.12	62.7
CMB	29.85	67.42	82.58	86.97	89.09	71.18

TABLE II
SID ACCURACY (%) OF THE PROPOSED METHODS WITH THE IBM

Babble	-6dB	0dB	6dB	12dB	18dB	Avg.
REC	27.58	50	71.06	81.21	88.94	63.76
MAR	69.85	77.73	85	88.48	90.30	82.27
CMB	60.91	79.70	87.27	91.97	93.33	82.64
Factory	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
REC	21.97	47.42	68.94	81.06	87.27	61.33
MAR	51.97	70.30	80.45	86.06	87.58	75.27
CMB	46.36	73.03	85	91.52	92.73	77.73
SSN	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
REC	19.70	43.94	66.21	79.39	86.36	59.12
MAR	55.15	74.70	82.27	87.73	89.24	77.82
CMB	52.27	75.30	85.61	91.36	93.18	79.54

TABLE III
SID ACCURACY (%) OF THE COMBINED SYSTEM AND BASELINES. PERFORMANCE IS AVERAGED ACROSS DIFFERENT SNR CONDITIONS

Method	Babble	Factory	SSN	Average
Combined System	72.58	71.33	71.18	71.7
GFCC_22	47.64	51.46	49.61	49.57
MFCC_22	39.42	35.95	31.58	35.65
MFCC_12	35.27	29.7	26.55	30.51
ETSI-AFE_D	40.55	45.33	43.27	43.05

in bounded marginalization [see (9)] takes time. These factors lead to the reconstruction module taking about 1/3 of the computing time of the marginalization module.

The combined system attempts to take advantage of the two methods. By looking at the SID results in Table I, it is clear that on average the marginalization module works better than reconstruction module for babble and SSN. The combined system significantly outperforms the individual modules.

To evaluate the quality of IBM estimation, we present the SID performance using the IBM in Table II. The table shows that both modules work very well using the IBM, especially the marginalization module. Compared with Table I, the reconstruction module has less significant improvement than the marginalization module. This may reflect the robustness of the reconstruction module to mask estimation errors. The dramatic gap between the two modules at -6 dB leads to a little performance degradation in the combined system. However, at 0 dB, although the gap is still large, the combined system is able to further improve the individual results.

Equation (11) weights the two modules equally. This combination is very simple, and it is possible that using unequal weights, e.g., assigning a higher weight to the more accurate module, produces better identification results. In the above IBM evaluation, we have found that, when we weight the two modules proportional to the numbers of selected frames, the performance of the combined system is improved a little compared to (11) as marginalization uses more active frames and therefore contributes more to the combination.

Table III lists the average SID results of the combined system along with those of the baseline systems given in Section VII-C. Clearly the combined system outperforms all three baselines. The combined system's SID results are more than 28 percentage points higher than those of MFCC and ETSI-AFE_D baselines. The gain over the GFCC baseline is smaller, reflecting the robustness of GFCC features themselves.

Under clean conditions, MFCC_22 yields the SID accuracy of 96.67% (94.39% for MFCC_12), whereas the accuracy is 97.12% for GFCC_22. GF as a spectral feature gives the accuracy of 95.76%, which is slightly worse than the 22-dimensional cepstral features. In a similar task on the 2002 NIST dataset, the accuracy of 89.39% was reported on the clean test set using MFCC features [1].

E. Comparison With a Related System

Pullella *et al.* recently proposed a system for robust speaker recognition, which also utilizes bounded marginalization to achieve noise robustness [23]. The difference from our marginalization module lies in two aspects. First, we use the gammatone filterbank as the front-end followed by decimation to derive GF features. They use a mel-scale filterbank as the front-end. The second difference is in mask estimation. They compute a binary mask using spectral subtraction, and then feature selection to refine the initial mask. It is questionable whether spectral subtraction can effectively deal with nonstationary noises. As described earlier, our system uses CASA-based speech segregation to directly estimate the IBM.

Our comparison uses the same experimental setup as in [23]. The speech signals are from the TIDigits corpus [14], from which 31 speakers (21 males and 10 females) are randomly chosen. Each speaker has speech utterances corresponding to 77 connected digits. Out of them, 50 are randomly chosen for training and 27 for testing. Test utterances are corrupted by white noise and factory noise at -5, 0, 5, 10, 15, and 20 dB. In the following figures, the performance of their system and MFCC baseline is directly taken from [23]. It is worth

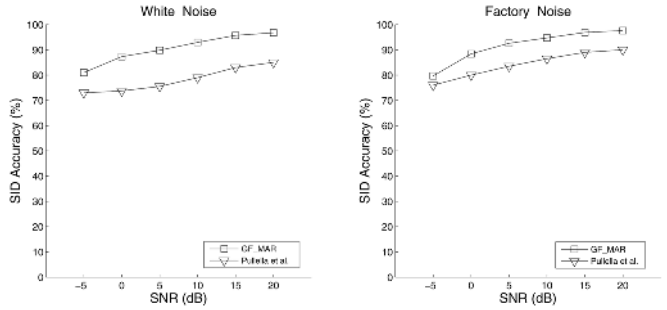

Fig. 6. SID accuracy (%) comparisons of the proposed marginalization module and Pullella *et al.*'s system. Both systems utilize the ideal binary mask for separation.

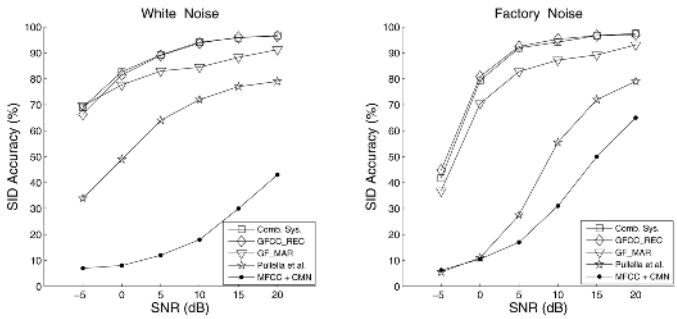

Fig. 7. SID accuracy (%) comparisons of the proposed combined system and Pullella *et al.*'s system using estimated binary masks.

mentioning that in this simulation we use individually trained GMMs instead of the GMM-UBM scheme to be consistent with their system, and the frame selection step is not employed due to relative short test utterances. The mask estimation process is the same as described in Section VII-A except that babble, white, factory and destroyer (operation room) noises are employed for MLP training.

Fig. 6 shows the SID performance with the IBM. To sharpen the comparison, we give the performance of the marginalization module of our system. The figure shows that our marginalization module yields SID accuracies that are about 10 percentage points higher than those in [23] for both white noise and factory noise conditions. In our system, only active frames are used for recognition, while their system appears to use all the frames. In this case, our system using all the frames achieves almost the same performance as using active frames. Therefore, this improvement should reflect the relative advantage of GF features over their mel-scale features.

Fig. 7 shows the SID performances of the proposed methods and Pullella *et al.*'s system with their respective methods of mask estimation. The comparison shows that all of our proposed methods perform much better than their system in both noise conditions, particularly at lower SNR levels. While our methods' performance does not vary a lot for the two noises, their system performs considerably worse in the factory noise, presumably because of the ineffectiveness of spectral subtraction for attenuating this nonstationary noise.

Comparing Figs. 6 and 7, our system with estimated binary masks does not degrade the performance by much compared to the use of the IBM, unlike the performance gaps in the NIST corpus shown earlier (cf. Tables I and II). We believe that this can be attributed to the large lexicon overlap between training and testing in the TIDigits corpus, which has a very small vocabulary. In the NIST corpus, there is no overlap between training and test utterances. We will come back to this point in the next section.

VIII. Discussion

An important finding in our study is that GFCC features outperform conventional MFCC features under noisy conditions. MFCC is obtained by a discrete Fourier transform (DFT), followed by a conversion to the Mel-frequency scale with a bank of triangular filters. Applying DCT to the log energy of the filter output produces MFCC. There are two main differences between GFCC and MFCC. First, GFCC uses a gammatone filterbank whereas MFCC uses a triangular filterbank applied to DFT. Gammatone filters constitute a more accurate model of cochlear filtering than triangular filters. Second, a log operation is applied in deriving MFCC whereas a cubic root operation is used in GFCC derivation. We believe that the performance advantage of GFCC is mainly due to the first difference, which is corroborated by the comparison in Fig. 6.

Our earlier work used the speech separation and recognition corpus (SSC) [5] as our test data [29], [30], and achieved large performance gains (see also [15]). However, we have found that such gains are somewhat inflated by the large lexicon overlap between training and test material. The SSC corpus has a small vocabulary and a large amount of training data. Each sentence in SSC has a fixed grammar and every word appears in both training and testing data. This situation is similar to the TIDigits corpus discussed in Section VII-E. On the other hand, the NIST corpus is a standard dataset for speaker recognition, which is much closer to practical situations.

Our previous work also employed uncertainty decoding in conjunction with GF feature reconstruction [29], [30]. Theoretically, uncertainty decoding is expected to improve recognition performance as the contributions of unreliable feature dimensions are discounted during decoding. Our experiments show that ideal information about feature uncertainty can indeed bring about considerable performance improvement. However, with estimated uncertainty, the decoding process does not provide significant performance improvements due to inevitable errors in the estimation process. How to estimate spectral uncertainty accurately is an interesting topic for future research.

In robust speech recognition, the reconstruction method shows better performance compared with bounded marginalization in larger vocabulary tasks [24], [33]. In our SID results, marginalization generally produces better results than reconstruction. The effectiveness of marginalization for SID has been shown in a number of previous studies [7], [23], [31]. We should note that speaker and speech recognition are two different tasks despite the fact that approaches are often shared between them.

Although the combined system in this study significantly outperforms the individual modules on the NIST dataset, the improvement on the Tidigits dataset is insignificant. The simple combination strategy in (11) seems to lose its advantage when the performance profiles of the individual modules are similar. In such situations, more sophisticated methods of classifier combination may be needed. Our future work will investigate this topic, which is a promising direction for further progress.

We should point out that this paper deals with additive noise in robust speaker identification, not handset/channel variations which are widely studied topics in robust speaker recognition. Are GF and GFCC features also robust to handset variations? There is no reason to believe so as these features are not designed for such variations. Whether common techniques for handling convolutive distortions such as CMN can be effectively combined with our approach to deal with both additive noise and handset variations is an interesting topic for future research.

To conclude, we have proposed new methods for robust speaker identification in noisy conditions, including novel speaker features of GF and GFCC. By using CASA masks for speech segregation, we can either reconstruct or marginalize unreliable components. Our systematic evaluations show that the proposed systems and their combination achieve significant performance improvements over alternative SID systems.

ACKNOWLEDGMENT

The authors would like to thank G. Hu and Z. Jin for their very valuable help in speech segregation, and the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] V. R. Apsingekar and P. L. De Leon, "Speaker model clustering for efficient speaker identification in large population applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 848–853, May 2009.

[2] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[3] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[4] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.

[5] M. Cooke and T. Lee, "Speech separation and recognition competition," 2006 [Online]. Available: http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm

[6] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[7] M. El-Maliki and A. Drygajlo, "Missing features detection and handling for robust speaker verification," in *Proc. Eurospeech*, 1999, pp. 975–978.

[8] S. Furui, "40 years of progress in automatic speaker recognition," *Lecture Notes Comput. Sci.*, vol. 5558, pp. 1050–1059, 2009.

[9] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[10] Y. Gong, "Noise-robust open-set speaker recognition using noise-dependent Gaussian mixture classifier," in *Proc. ICASSP*, 2002, pp. 133–136.

[11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[12] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1091–1102, Jul. 2011.

[13] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, May 2009.

[14] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, 1984, pp. 328–331.

[15] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *Proc. ICASSP*, 2010, pp. 4514–4517.

[16] Y. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Commun.*, vol. 51, pp. 230–239, 2009.

[17] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Comput. Speech Lang.*, vol. 10, pp. 107–116, 1996.

[18] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. San Diego, CA: Academic, 2003.

[19] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[20] N. Parihar and J. Picone, "Analysis of the Aurora large vocabulary evaluations," in *Proc. Eurospeech*, 2003, pp. 337–340.

[21] R. D. Patterson, J. Holdsworth, and M. Allerhand, "Auditory models as preprocessors for speech recognition," in *The Auditory Processing of Speech: From Sounds to Words*, M. E. H. Schouten, Ed. Berlin, Germany: Mouton de Gruyter, 1992, pp. 67–83.

[22] M. Przybocki and A. Martin, "The NIST Year 2002 Speaker Recognition Evaluation Plan," 2002 [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2002/2002-spkrec-evalplan-v60.pdf

[23] D. Pullella, M. Kühne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Proc. ICASSP*, 2008, pp. 4833–4836.

[24] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, pp. 275–296, 2004.

[25] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech*, 1997, pp. 963–966.

[26] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, 1995.

[27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[28] A. Schmidt-Nielsen and T. H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Process.*, vol. 10, pp. 249–266, 2000.

[29] Y. Shao, S. Srinivasan, and D. L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. ICASSP*, 2007, pp. 277–280.

[30] Y. Shao and D. L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. ICASSP*, 2008, pp. 1589–1592.

[31] Y. Shao and D. L. Wang, "Robust speaker recognition using binary time-frequency masks," in *Proc. ICASSP*, 2006, pp. 645–648.

[32] E. Shriberg, "Higher-level features in speaker recognition," *Lecture Notes Comput. Sci.*, vol. 4343, pp. 241–259, 2007.

[33] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1486–1501, 2006.

[34] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2130–2140, Sep. 2007.

[35] *ETSI Standard, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms"*, ETSI ES 202 050 v1.1.4, 2005, European Telecommunications Standards Institute, ETSI ES 202 050 v1.1.4.

[36] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.

[37] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE, 2006.

[38] N. Wang, P. C. Ching, N. Zheng, and T. Lee, "Robust speaker recognition using denoised vocal source and vocal tract features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 196–205, Jan. 2011.

[39] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proc. ICASSP*, 2001, pp. 457–460.

Xiaojia Zhao (S'11) received the B.E. degree in software engineering from Nankai University, Tianjin, China, in 2008. He is currently pursuing the Ph.D. degree at The Ohio State University, Columbus.

His research interests include computational auditory scene analysis, speaker/speech recognition, and statistical machine learning.

Yang Shao, photograph and biography not available at the time of publication.

DeLiang Wang, (F'04) photograph and biography not available at the time of publication.