Robust Speaker Identification in Noisy and Reverberant Conditions

Xiaojia Zhao, Yuxuan Wang, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Robustness of speaker recognition systems is crucial for real-world applications, which typically contain both additive noise and room reverberation. However, the combined effects of additive noise and convolutive reverberation have been rarely studied in speaker identification (SID). This paper addresses this issue in two phases. We first remove background noise through binary masking using a deep neural network classifier. Then we perform robust SID with speaker models trained in selected reverberant conditions, on the basis of bounded marginalization and direct masking. Evaluation results show that the proposed system substantially improves SID performance over related systems in a wide range of reverberation time and signal-to-noise ratios.**

***Index Terms*—Deep neural network, ideal binary mask, noise, reverberation, robust speaker identification.**

I. Introduction

ROBUSTNESS of automatic speaker recognition is critical for real-world applications. In daily acoustic environments, additive noise, room reverberation and channel/handset variations conspire to pose considerable challenges to such systems. A lot of research has been devoted to dealing with individual challenges. For example, speakers can be modeled in multiple noisy environments to reduce the mismatch between training and test conditions [26]. Speech enhancement methods, such as spectral subtraction, have been explored for noise-robust speaker recognition [38]. Computational auditory scene analysis (CASA) was recently employed to remove noise [43]. Speaker features such as modulation spectral features [6] and those incorporating phase information [41] have shown robustness against reverberation. Blind dereverberation algorithms have been used to restore the anechoic signal or the early reflections of reverberant speech [33]. Borgstrom and McCree modeled the effect of reverberation as a channel-wise convolution of short-time spectral envelopes [3]. In this study, the room impulse response (RIR) is characterized as a causal low-pass filter in the modulation envelope domain, and linear prediction inverse modulation transfer function is estimated to remove the effect of reverberation. Alternatively, one can deliberately introduce reverberation to speaker models to reduce the mismatch caused by reverberation [1]. By and large, the speaker recognition community has focused on channel variations in speaker verification. The National Institute of Standards and Technology (NIST) has conducted a series of speaker recognition evaluations (SRE) since 1996. State-of-the-art systems include joint factor analysis [18] and i-vector based techniques [5].

However, efforts have rarely been made on the combined effects of noise and reverberation. May *et al.* [24] and Gonzalez-Rodriguez *et al.* [9] studied the combined effects using binaural cues and microphone arrays. Garcia-Romero *et al.* [7] and Krishnamoorthy and Prasanna [20] reported results in noisy and reverberant conditions separately but not together. It is worth noting that studies on human listeners suggest the combined effects of noise and reverberation degrade speech intelligibility to a greater degree than individually [13], [27].

In this study, we explore the combined effects of noise and reverberation in monaural speaker identification (SID). We deal with reverberation by training models in noise-free reverberant conditions, while assuming little knowledge of the amount of reverberation in the test data. Meanwhile, noise is suppressed through a CASA approach that segregates speech by binary time-frequency (T-F) masking. We perform binary classification using a deep neural network (DNN). We utilize a CASA mask for SID in two ways, namely bounded marginalization and direct masking. The outputs of the two methods are combined to make the final SID decision.

The rest of the paper is organized as follows. Section II gives an overview of the system and discusses front-end processing including DNN-based mask estimation. Bounded marginalization and direct masking are introduced in Section III, followed by evaluations in Section IV. We conclude this paper in Section V.

II. System overview and Front-end Processing

Fig. 1 shows the schematic diagram of the proposed system. Noisy speech is first passed through a DNN classifier to produce a binary T-F mask. Simultaneously we extract*gammatone features* (GF) and *gammatone frequency cepstral coefficients* (GFCC) [34]. Each of the multiple training conditions produces one set of speaker models that is utilized independently. GF-based speaker models are fed to the bounded marginalization module, while GFCC-based speaker models to the direct masking module. Local decisions corresponding to different training conditions are first combined within each module

Manuscript received April 16, 2013; revised August 04, 2013; accepted February 14, 2014. Date of publication February 25, 2014; date of current version March 06, 2014. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-12-1-0130. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rodrigo Capobianco Guido.

X. Zhao and Y. Wang 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; wangyuxu@cse.ohio-state.edu).

D. Wang is with the Department of Computer Science and Engineering and Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210-1277 USA.

Digital Object Identifier 10.1109/TASLP.2014.2308398

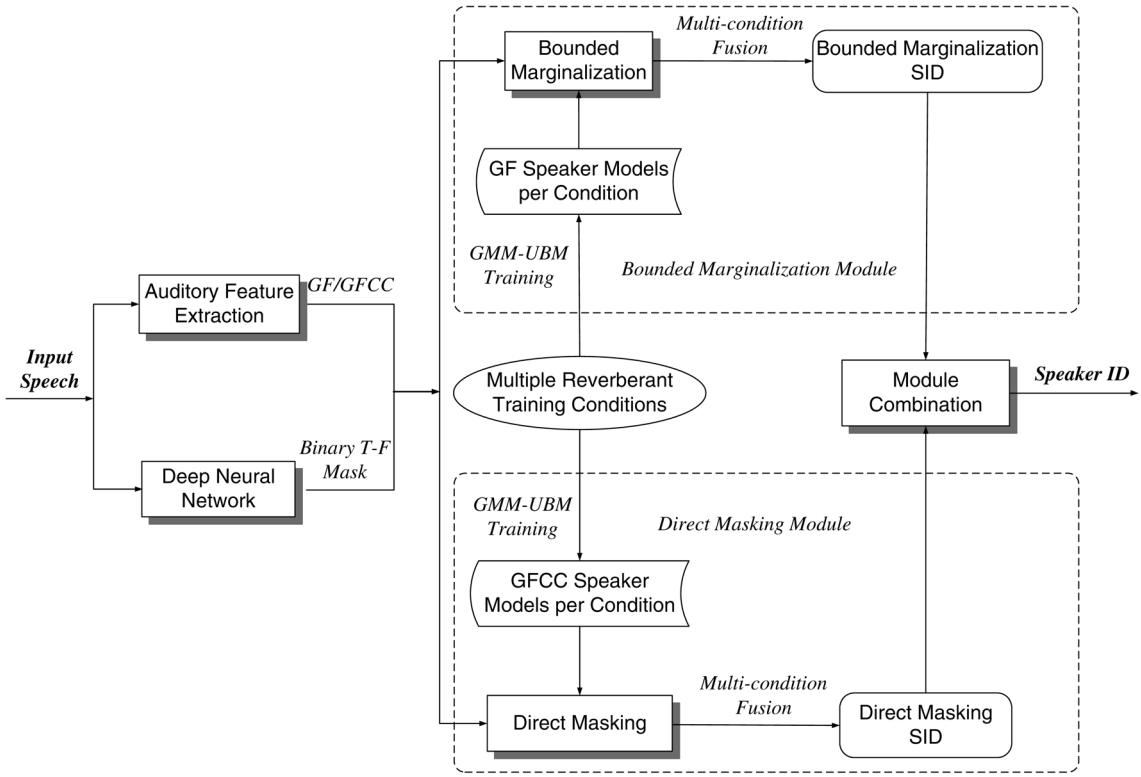

Fig. 1. Schematic diagram of the proposed speaker identification system.

and subsequently between two modules to make the final SID decision. Below, we describe auditory features and discuss the different definitions of a CASA mask in the noisy and reverberant conditions. Then DNN-based binary masking is described.

A. Auditory Features

Two auditory features are employed in our system. One is GF in the spectral domain and the other one is GFCC in the cepstral domain. They are chosen primarily because of their robustness relative to other commonly used speaker features such as *mel-frequency cepstral coefficients* (MFCC) [43].

Noisy and reverberant speech is first passed through a 64-channel gammatone filterbank to create a two-dimensional *cochleagram* [37]. Each frame of the cochleagram is rectified using the cubic root operation to generate a GF vector. We apply discrete cosine transform to GF to derive GFCC. Detailed feature extraction can be found in [43].

B. Mask Definitions in Noisy and Reverberant Conditions

A main computational goal of CASA is the *ideal binary mask* (IBM), where each element corresponds to a T-F unit in the cochleagram and indicates whether the corresponding unit is dominated by target or interference [36]. In this paper, the *target* refers to the speech signal, and the *interference* the background of the target speech. The IBM is defined as follows:

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

$SNR(t,f)$ refers to the local signal-to-noise ratio (SNR) of the T-F unit in time frame t and frequency channel f. LC denotes an SNR threshold called *local criterion*. Given premixed target and interference signals, the IBM can be readily constructed. The IBM concept is motivated by the auditory masking phenomenon and is the optimal binary mask in terms of SNR gain [22].

What constitutes the target signal is not a straightforward question in noisy and reverberant conditions. For example, the entire reverberant speech can be considered as the target and the reverberant noise as interference [16]. We call this the *Reverberant IBM* ($\mathrm{IBM_R}$). Meanwhile, one can choose only the early reflections of the reverberant speech as the target and everything else (i.e. late reverberation and reverberant noise) as interference [32]. The resulting definition is named *Early-Reverberant IBM* ($\mathrm{IBM_{ER}}$). If we treat only the direct path of the reverberant speech as the target, we can obtain *Direct-Sound IBM* ($\mathrm{IBM_{DS}}$) [23]. We explore these three IBMs in Section IV.

C. Mask Estimation via DNN

The definition of the IBM is based on the prior information of target and interference. In practice, we have to estimate the IBM. Recent work in CASA employs supervised classification for IBM estimation. Gaussian mixture models (GMMs) [19] and support vector machines (SVMs) [11] have been used in anechoic conditions. Motivated by their superior performance [39], we employ DNNs for mask estimation in this study. The employed mask estimation system is detailed below.

We use the standard generative-discriminative procedure to train DNNs. First, the DNNs are pretrained using restricted Boltzmann machines (RBMs) in an unsupervised and layerwise fashion. An RBM is a two-layer neural network with a visible layer v and hidden layer h, and a stack of RBMs forms a very

powerful generative model [15]. The joint probability of an RBM is defined based on an energy function E:

$$p(v,h) = \frac{e^{-E(v,h)}}{Z}, \quad (2)$$

where Z is a normalization term called partition function. More specifically, raw features are used to train the first RBM. We then take the hidden activations from the first RBM to train the second RBM, and so on. Since the inputs (raw features) are usually real valued, we employ a Gaussian-Bernoulli RBM for the first layer and Bernoulli-Bernoulli RBMs for all subsequent layers. Assuming visible units are Gaussian random variables with unit variance, we can define the energy function E for this Gaussian-Bernoulli RBM as:

$$E(v,h) = \frac{1}{2}\sum_i (v_i - a_i)^2 - \sum_j b_j h_j - \sum_{i,j} w_{ij} v_i h_j, \quad (3)$$

where v_i and h_j are the ith and jth unit of v and h, a_i and b_j are the bias for v_i and h_j, respectively, and w_{ij} is the symmetric weight between v_i and h_j.

Training an RBM requires maximizing the joint probability (2) with respect to network weights. Once pretrained, the weights from a stack of RBMs are used to initialize a standard feedforward network, which is then discriminatively fined-tuned using the backpropagation algorithm. Since our target labels are binary, we use the cross-entropy objective function for backpropagation:

$$E(v,h) = \sum_m \left[d^{(m)} \log(p^{(m)}) + \left(1 - d^{(m)}\right) \log\left(1 - p^{(m)}\right)\right], \quad (4)$$

where m indexes training samples, $d^{(m)}$ is the label of sample m and $p^{(m)}$ is the corresponding network prediction (posterior probability).

Our separation system works as follows. We extract features from the cochleagram and train a subband classifier for each frequency channel to estimate the target-dominance of each T-F unit, where the training labels are provided by the IBM. Since a decision needs to be made for each T-F unit, we extract unit-level features from the subband signal within each T-F unit. In this study, we use the complementary feature set proposed in [40], which consists of amplitude modulation spectrogram, RASTA-PLP, MFCC and pitch-based features. We used the DNNs described above as the subband classifiers.

III. Recognition Methodology and System Design

Over the past few decades, the GMM has been the predominant approach for speaker modeling [30]. The GMM framework along with the universal background model (UBM) [31] is adopted for speaker modeling in this study. The feature space of a speaker is described as a linear combination of multivariate Gaussians that represent broad acoustic classes. Such Gaussians are usually parameterized with diagonal covariance matrices. Given speaker models, we employ different recognition methods by incorporating binary masking. At each frame, a binary mask divides the T-F units into two groups. One group consists of reliable T-F units with the label of 1 while the remaining unreliable T-F units, with the label of 0, form the other group. Multiple methods have been developed to deal with unreliable T-F units group such as marginalization, reconstruction, and direct masking. We use bounded marginalization and direct masking as two modules.

A. Bounded Marginalization Module

The basic idea of marginalization is to base recognition on reliable T-F units while removing the impact of unreliable ones. Conventional marginalization integrates over unreliable T-F units in the entire range of feature values, e.g. minus infinity to positive infinity. Bounded marginalization sets realistic lower and upper bounds for the integration, which has proven beneficial [25], [43]. Its analytical form is written as follows [4], [43].

$$\begin{aligned} L(X\,|\lambda) &= \int_{low}^{high} p(X\ |\lambda) dX_u \\ &= \int_{low}^{high} p(X_r, X_u\,|\lambda) dX_u \\ &= \int_{low}^{high} \sum_{k=1}^{K} p(k) p(X_r, X_u\,|k) dX_u \\ &= \sum_{k=1}^{K} p(k) p(X_r\,|k) \int_{low}^{high} p(X_u\,|k) dX_u. \end{aligned} \quad (5)$$

Here X_r denotes the set of reliable T-F units and X_u the set of unreliable ones. The likelihood $L(X|\lambda)$ of a frame vector X produced by a speaker λ can be calculated by integrating the probability density function of λ with respect to X_u, from the lower bound (i.e. *low*) to the upper bound (i.e. *high*). The integration can be carried out in each of the K Gaussians of the GMM. Specifically, we perform bounded marginalization on the GF features with a CASA mask specifying the reliable and unreliable T-F units.

B. Direct Masking Module

Direct masking is a recently proposed technique for coupling binary masking and speech recognition [12]. In direct masking, one simply attenuates the noise-dominant T-F units using a constant gain, instead of estimating them as done in feature reconstruction. Cepstral features are then calculated directly from this masked representation or from the resynthesized target signal. Results have shown that this leads to competitive recognition performance compared to bounded marginalization and feature reconstruction. Therefore, we use direct masking in this study.

When the IBM is available, we retain target-dominant T-F units and attenuate noise-dominant T-F units by 26 dB. For estimated binary masks, we have found that using the outputs of the DNNs directly performs better than converting them to binary values. GFCC features for speaker recognition are extracted from the resynthesized target signal, which is obtained by applying the ratio mask (i.e. DNN output) to the mixture.

C. Reverberant Model Training

Speaker models trained in anechoic and noise-free conditions do not generalize well to reverberation. To characterize speaker

feature distributions in such conditions, we train speaker models from reverberant environments.

Reverberation is usually characterized in terms of *reverberation time* (T_{60}), which describes the amount of time for the direct sound to decrease by 60 dB. Room reverberation is typically modeled as a convolution between a direct signal and an RIR which characterizes a specific reverberant condition. An RIR is determined by many factors such as geometry of the room, locations of sound sources and receivers.

To simplify the experimental settings while assuming little prior knowledge of testing reverberation, we simulate N reverberant environments covering a plausible range of T_{60}. In this study, the range is chosen from 0s (anechoic condition) up to 1s, covering daily room environments [21]. These N reverberant conditions are chosen as the representatives of the range and expected" to generalize to T_{60}'s between these representative values. We train a set of speaker models in each of the N conditions. Each set of speaker models characterizes a unique reverberant condition and is used independently for speaker recognition.

D. Multi-condition Fusion and Module Combination

In each of the two modules, SID decisions from the N sets of speaker models are first fused to generate module output. We then combine the outputs of the two modules to make the final SID decision. This design is elaborated below.

For an unknown test reverberant condition, each of the N reverberant training conditions correlates with the test condition differently. The speaker models from the best matching conditions should be used. However, these correlations are unknown without ground truth information. Reverberation classification has been proposed to classify the test reverberant condition as one of the training conditions and select speaker models from the chosen condition for recognition [1], [29]. There are two problems with this idea. The first one occurs when the test condition does not match any of the training conditions. A hard classification is unlikely to work well. The second is that the idea was tested only in noise-free reverberant conditions. It is more challenging to perform such a classification task when background noise is present. Instead of reverberation classification, we propose to fuse the contributions from all training conditions. If done well, we expect that the best matching condition will dominate the fusion. If none of the training conditions match the test condition well, this fusion could leverage multiple contributions. As the score ranges from these conditions could be very different, we normalize before fusing them to make the final SID decision. The normalization is described in the following equation,

$$\hat{s} = \frac{s - \min(s)}{\max(s) - \min(s)}. \tag{6}$$

where s refers to the output of a single training condition, which is a vector of scores with the number of elements equal to the number of speakers used in training. $\hat{s}$ denotes the normalized score vector. We combine the N normalized score vectors using a simple summation. We have also explored several other combination strategies and none of them significantly outperforms this simple summation. The fusion is performed in both bounded marginalization and direct masking modules.

The two modules address SID in noise from different perspectives. The bounded marginalization module works in the spectral domain and utilizes some information from unreliable T-F units. On the other hand, the direct masking attenuates unreliable T-F units uniformly and employs GFCC in the cepstral domain. GF and GFCC exhibit complementary properties for noise-robust SID [43]. We have observed that the errors of the two modules tend not to agree and the underlying speaker often achieves high scores in both modules. Hence, we combine these two modules to further improve SID performance. Similar to within-module fusion, we first apply score normalization (see (6)) and then simply add the module scores.

IV. Evaluation and Comparison

A. Experimental Setup

We randomly drew 300 speakers from the 2008 NIST Speaker Recognition Evaluation dataset (*short 2* part of the training set). Each speaker has a telephone conversation excerpt of 5 minutes in total duration. We apply simple energy-based voice activity detection to remove the large chunks of silence in the excerpt. Then we divide the recording into 5s long pieces. Two pieces with the highest energy are selected as the test data in order to provide sufficient speech information. The remaining pieces are used for training. On average there are about 20 training utterances per speaker. We employ the Matlab implementation of the image method of Allen and Berkley [2] to simulate room reverberation [10]; results with recorded impulse responses are given in Section IV.E. The range of T_{60} is varied from 0 to 1s, which covers a broad range of realistic reverberant environments [21]. We simulate three rectangular rooms to obtain 3 T_{60}'s: 300, 600 and 900 ms. For each T_{60}, we simulate 5 RIRs by randomly positioning a speech source and a receiver with the source-to-receiver distance fixed at 2 m. Each training utterance is convolved with the 5 RIRs. Each speaker is modeled in these three T_{60}'s separately using the GMM-UBM framework [31]. Test RIRs, on the other hand, are obtained from 7 simulated rooms corresponding to 7 T_{60}'s from 300 ms to 900 ms with the increment of 100 ms. Details of the simulated rooms are shown in Table I. We simulate 3 pairs of RIRs in each room (T_{60}) by randomly positioning a speech source, a noise source and a receiver with both source-to-receiver distances fixed at 2 m. The relative location of each source to the receiver determines an RIR. This results in 21 pairs of RIRs in total. Each test utterance is convolved with 2 pairs of RIRs that are randomly selected from the 21 pairs RIR library. Specifically, for each pair, the RIR corresponding to the speech source is used to convolve with the target speech and the other one with interference. Factory noise, speech shape noise (SSN) and destroyer engine room noise from the Noisex-92 database are used as interference [35]. We generate 5 SNRs for each noise from 0 to 24 dB with the increment of 6 dB. In total, each SNR of each noise has $300^*2^*2 = 1200$ test trials.

We use two-hidden-layer DNNs, which strike a balance between performance and computational overhead [39]. We

TABLE I
SIZES OF THE SIMULATED RECTANGULAR ROOMS

T_{60} (ms)	Training Rooms: length, width, height (m)	Testing Rooms: length, width, height (m)
300	(5, 4, 3)	(5, 4, 3)
400	-	(6, 4, 3)
500	-	(7, 5, 4)
600	(7, 6, 4)	(7, 6, 4)
700	-	(8, 7, 5)
800	-	(8, 7, 6)
900	(9, 8, 7)	(9, 8, 7)

train DNNs separately for bounded marginalization and direct masking. The IBM is used to provide training labels. The IBM compares the local SNR of each T-F unit with the LC, which is typically set to 0 dB to indicate which source is stronger. Recent studies on speech intelligibility [32] and robust speech recognition [28], however, have shown that an LC of 0 dB is not always optimal, which is confirmed by our SID experiments. We will elaborate how LC is determined in the following subsection. We also compare the three reverberant IBM definitions in terms of SID performance. The DNN training set is created by mixing 50 utterances from 50 randomly selected speakers with the 3 noises, 3 training T_{60}'s (300, 600, and 900 ms), and 5 SNR conditions (-5, 0, 5, 10 and 15 dB). At each T_{60}, the 5 training RIRs are divided into two groups: three are used to create the DNN training set and the remaining two for a cross validation set. In other words, the 50 DNN training utterances are convolved with 3 selected RIRs. The remaining 2 RIRs are used to distort another 10 randomly selected training utterances to create a validation set for DNN training. For each RIR, we randomly choose a noise source position within the same room to derive an RIR for the noise. This gives us noisy and reverberant DNN training and validation sets. Note that the RIRs used for speaker modeling as well as DNN training/testing do not overlap with the RIRs of SID evaluations. The DNNs are supervisedly fine-tuned using stochastic gradient descent with the objective function given in (4).

We extract 64-dimensional GF for bounded marginalization and 22-dimensional GFCC features for direct masking. We also extract 22-dimensional MFCC features for the sake of comparison. Speaker models are adapted from a 1024-component UBM that is trained by pooling training data from all the enrolled speakers [31]. For each speaker, we train 3 sets of models in the three reverberant training conditions for GF, GFCC and MFCC respectively. In addition, we train a set of anechoic models for each feature to generate benchmark performance.

We perform SID in selected frames with some target information. We refer to the frames containing at least one reliable T-F unit as "active frames". To balance the number of selected frames and the number of reliable T-F units per frame to qualify for selection, we use as the selection criterion the smaller of half of the number of all frequency channels (i.e. 32) and the median number of reliable T-F units of all active frames for a noisy and reverberant speech utterance. Given an active frame, it will be selected if its number of reliable units is greater than the criterion.

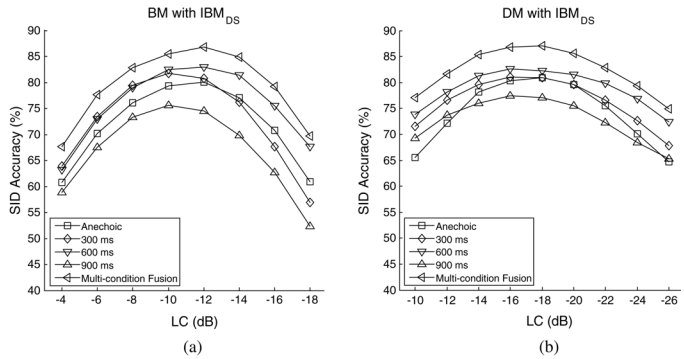

Fig. 2. SID accuracy (%) using $\mathrm{IBM}_{\mathrm{DS}}$ with different LCs. **BM** denotes bounded marginalization, **DM** direct masking. (a) BM performance. (b) DM performance. Each point in the figure is averaged across all the test SNRs. **Anechoic** denotes speaker models trained in the anechoic condition. **300 ms**, **600 ms** and **900 ms** denote speaker models trained in the corresponding T_{60}. **Multi-condition Fusion** combines the local decisions from the four sets of speaker models.

TABLE II
OPTIMAL LCS (DB) FOR DIFFERENT IBM DEFINITIONS

IBM Definition	Bounded Marginalization	Direct Masking
$\mathrm{IBM}_{\mathrm{R}}$	-4	-12
$\mathrm{IBM}_{\mathrm{ER}}$	-4	-12
$\mathrm{IBM}_{\mathrm{DS}}$	-12	-18

B. IBM Comparisons

Regarding different IBM definitions as discussed in Section II.B, one important issue is which IBM is most effective for SID. A related issue is LC values in the IBM definition.

To address these issues, we set up a small experiment by randomly selecting 50 speakers from the TIMIT corpus. Each speaker has 10 utterances in total, 8 of which are used for training and the remaining 2 for testing. Training and testing data are mixed with newly simulated RIRs following the same procedure as described earlier except that the sampling frequency is 16000 Hz (8000 Hz for the NIST SRE dataset). Only factory noise is used in this experiment and the SNRs of the test set are -6, 0, 6, 12 and 18 dB. Ten RIRs are randomly selected from the 21 test RIR pairs, so there are $50^{*}2^{*}10 = 1000$ test trials for each SNR condition.

Fig. 2 gives an example of $\mathrm{IBM}_{\mathrm{DS}}$. We vary LC to generate different IBMs for the two modules separately. Results are averaged across the 5 SNRs. The figure indicates that an LC of -12 dB is the optimal choice for bounded marginalization. On the other hand, direct masking favors an LC of -18 dB. Note that the plateau of each plot is relatively wide and it shows that the proposed SID system is robust to LC choices. It is worth mentioning that the proposed multi-condition fusion idea works well as expected. We conduct similar experiments on the other two IBM definitions. Our obtained optimal LCs are listed in Table II.

An example of the three IBM definitions is shown in Fig. 3. The left plot in each panel was created with an LC of 0 dB. It retains a reasonable number of 1s in $\mathrm{IBM}_{\mathrm{R}}$. However, it gets

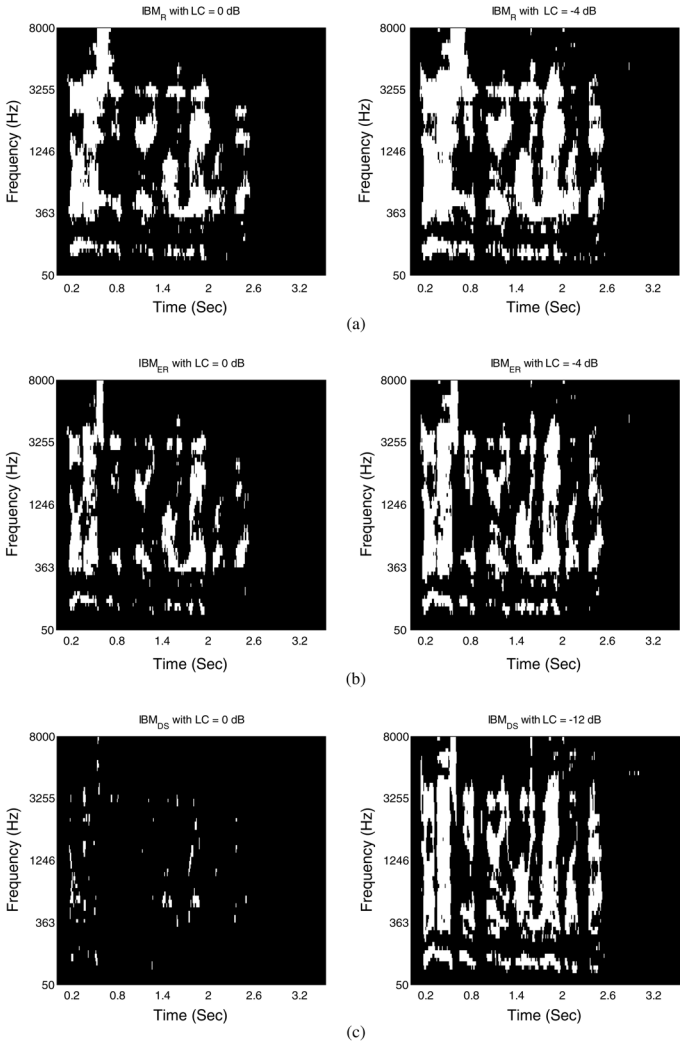

Fig. 3. Illustrations of 3 IBM definitions on a TIMIT sentence with 0 dB SNR and 500 ms T_{60}. (a) $\mathrm{IBM}_{\mathrm{R}}$ with two LCs, 0 dB on the left and -4 dB on the right. (b) $\mathrm{IBM}_{\mathrm{ER}}$ with two LCs, 0 dB on the left and -4 dB on the right. (c) $\mathrm{IBM}_{\mathrm{DS}}$ with two LCs, 0 dB on the left and -12 dB on the right. 1 is shown as white and 0 as black.

sparser for $\mathrm{IBM}_{\mathrm{ER}}$, and $\mathrm{IBM}_{\mathrm{DS}}$ has very few 1s. This is expected as $\mathrm{IBM}_{\mathrm{ER}}$ and $\mathrm{IBM}_{\mathrm{DS}}$ treat part or all the reverberated speech as interference, and their effective SNRs are much lower than $\mathrm{IBM}_{\mathrm{R}}$. As we choose the optimal LCs for bounded marginalization (the right plots), more 1s show up, and the three IBMs now exhibit similar patterns.

Next we explore the utilities of the three IBM definitions with their optimal LCs for SID. As shown in Table III, the proposed system outperforms the individual modules for all three IBM definitions. $\mathrm{IBM}_{\mathrm{DS}}$ produces the best performance in all the categories. The other two IBM definitions achieve comparable performance. It remains to be seen if the performance advantage of $\mathrm{IBM}_{\mathrm{DS}}$ holds when estimated IBMs are employed and a larger dataset like the NIST SRE is used.

C. Performance with Estimated IBM

We now establish benchmark SID performance of the NIST SRE dataset. We apply anechoic speaker models to the anechoic and all reverberant test sets where noise is excluded. As shown

TABLE III
SID ACCURACY (%) OF THE THREE IBM DEFINITIONS. PERFORMANCE IS AVERAGED CROSS ALL THE SNRS

IBM Definition	Bounded Marginalization	Direct Masking	Proposed System
$\mathrm{IBM}_{\mathrm{R}}$	83.68	83.18	88.52
$\mathrm{IBM}_{\mathrm{ER}}$	84.48	83.82	88.86
$\mathrm{IBM}_{\mathrm{DS}}$	86.78	87.02	92.32

TABLE IV
BENCHMARK SID PERFORMANCE (%) OF ANECHOIC SPEAKER MODELS. NOISE IS EXCLUDED IN THE TEST SET

Features	Anechoic Test Set	Reverberant Test Set
MFCC	97.83	77.08
GFCC	88.17	56.08
GF	95.00	54.42

TABLE V
BENCHMARK SID PERFORMANCE (%) OF REVERBERANT SPEAKER MODELS IN THE REVERBERANT TEST SET. NOTE THAT THE FIRST COLUMN IS THE SAME AS THE LAST COLUMN OF TABLE IV AS BOTH USE ANECHOIC SPEAKER MODELS TO RECOGNIZE REVERBERANT TEST DATA

Features	Anechoic Models	300 ms	600 ms	900 ms
MFCC	77.08	85.75	86.00	82.42
GFCC	56.08	75.17	77.33	73.92
GF	54.42	82.67	87.17	84.25

in Table IV, MFCC-based models achieve the best performance in both anechoic and reverberant conditions. When reverberation is included in the test set, the performance of all anechoic speaker models drops substantially due to the mismatch. After reverberation is included in speaker models, the performance is shown in Table V. As shown in the table, the introduction of reverberation in the training data significantly improves performance for all the features. The GF-based models even outperform MFCC-based models in some cases. Models trained in the T_{60} of 600 ms achieves the best performance, probably because it lies in the middle of the test T_{60} range.

Now we evaluate the proposed system in the noisy and reverberant test set. We use GF for the bounded marginalization module. Both GFCC and MFCC are used in the direct masking module. When estimated IBMs are used, we notice that the inclusion of anechoic speaker models in the multi-condition fusion stage does not help at all due to their substantial performance gap from reverberant speaker models. Therefore, we only fuse the reverberant speaker models in each module. Table VI shows the SID performances of the proposed methods: the direct masking module with MFCC and GFCC, the bounded marginalization module with GF, and the combined system. On average, the bounded marginalization module outperforms the direct masking module for both MFCC and GFCC. The direct masking module with GFCC substantially outperforms that of MFCC at the low SNRs, likely due to the better noise robustness of GFCC features [43]. As the SNR increases, MFCC closes the gap and even outperforms GFCC. In the combined system, we employ the direct masking module with GFCC. The combined system outperforms individual

TABLE VI
SID ACCURACY (%) OF THE PROPOSED SYSTEM USING $\mathrm{IBM_R}$. MFCC_DM DENOTES THE DIRECT MASKING MODULE WITH MFCC FEATURES. GFCC_DM DENOTES THE DIRECT MASKING MODULE WITH GFCC FEATURES. GF_BM DENOTES THE BOUNDED MARGINALIZATION MODULE. COMB. SYST. DENOTES THE PROPOSED SYSTEM

Factory	0dB	6dB	12dB	18dB	24dB	Average
MFCC_DM	12.50	23.00	43.50	64.83	75.67	43.90
GFCC_DM	33.25	48.33	61.00	70.33	74.00	57.38
GF_BM	34.08	49.17	59.25	71.92	80.00	58.88
Comb. Syst.	40.33	59.17	67.67	76.75	81.83	65.15

Destroyer	0dB	6dB	12dB	18dB	24dB	Average
MFCC_DM	16.17	33.17	50.33	64.17	75.17	47.80
GFCC_DM	35.75	47.25	57.67	66.58	72.08	55.87
GF_BM	45.83	57.92	69.08	78.58	81.83	66.65
Comb. Syst.	50.00	61.67	73.50	79.75	82.42	69.47

SSN	0dB	6dB	12dB	18dB	24dB	Average
MFCC_DM	18.50	34.33	55.92	71.33	79.58	51.93
GFCC_DM	37.67	52.92	64.33	70.92	74.92	60.15
GF_BM	44.83	61.17	73.58	79.83	83.42	68.57
Comb. Syst.	51.25	67.00	77.83	83.00	84.83	72.78

TABLE VII
SID PERFORMANCE (%) SUMMARY OF THE PROPOSED SYSTEM WITH DIFFERENT IBM DEFINITIONS

Factory	0dB	6dB	12dB	18dB	24dB	Average
$\mathrm{IBM_R}$	40.33	59.17	67.67	76.75	81.83	65.15
$\mathrm{IBM_{ER}}$	37.92	56.83	67.67	74.75	79.00	63.23
$\mathrm{IBM_{DS}}$	35.00	53.75	64.25	74.42	79.08	61.30

Destroyer	0dB	6dB	12dB	18dB	24dB	Average
$\mathrm{IBM_R}$	50.00	61.67	73.50	79.75	82.42	69.47
$\mathrm{IBM_{ER}}$	48.42	62.42	72.42	79.33	81.50	68.82
$\mathrm{IBM_{DS}}$	43.00	57.08	70.17	77.83	80.50	65.72

SSN	0dB	6dB	12dB	18dB	24dB	Average
$\mathrm{IBM_R}$	51.25	67.00	77.83	83.00	84.83	72.78
$\mathrm{IBM_{ER}}$	49.67	66.75	75.67	81.42	81.58	71.02
$\mathrm{IBM_{DS}}$	41.75	62.42	74.25	80.08	81.58	68.02

modules in every test condition. For example, the combined system outperforms both modules by around 10% for factory noise at 6 dB.

Table VII lists the SID results of the proposed system with the three IBM definitions. Compared to Table III, the advantage of $\mathrm{IBM_{DS}}$ no longer exists. It is likely due to the quality of estimated masks. $\mathrm{IBM_{DS}}$ has very low LCs (-12 and -18 dB), which introduce substantial amounts of noise to the reliable T-F units, making it difficult for mask estimation. The performances of the other two IBMs definitions are still close. These results suggest that it is easier to estimate $\mathrm{IBM_R}$.

D. Comparison with Related Systems

We pointed out that there was little study on the combined effects of reverberation and noise for SID. It is thus difficult to find comparison systems. As a result, we adapt a few related systems for the sake of comparison which should still provide useful perspectives on the relative performance of our model. The first related system, labeled as "Multi-conditional Training" in Fig. 4, was designed for robust speaker verification using i-vector based techniques [7]. Each speaker is modeled as a GMM adapted from the UBM. A supervector is obtained by concatenating the means of Gaussians, and Garcia-Romero *et al.* [7] map the supervector to a lower-dimensional factor named an i-vector (or identity vector) (see also [5]). This system focuses on how to train from multiple training conditions followed by a combination to deal with noise and reverberation. A top performing scheme trains Gaussian probabilistic linear discriminant analysis models in both reverberant and noisy conditions. We implement this scheme for comparison due to its effectiveness and simplicity. In their experiments, multi-conditional training data were created by adding 3 types of noise: babble, car and helicopter, at 0 dB, 6 dB, 10 dB and 20 dB SNRs. Additional training data were produced by convolving clean speech with simulated RIRs at 100 ms, 300 ms and 500 ms reverberation times. Totally there are 16 training conditions, including the anechoic condition. In the implementation of this method, we use 19-dimensional MFCC features and their delta features to be consistent with the comparison system. Speaker models are trained by pooling training data from not only the reverberant training conditions we use, but also anechoic and noisy conditions (factory noise, destroyer engine room noise and SSN) in a wide range of SNRs (0, 6, 12, 18 and 24 dB). The second system, labeled as "Reverb. Classification", was designed to deal with reverberation alone [1], [29]. It trains speaker models in multiple reverberant conditions separately. Given a test utterance, it first identifies the closest training condition and uses the models of that condition to perform speaker recognition, as detailed in Section 2.2 of [1] and Section 3.4 of [29]. Since it only deals with noise-free reverberant speech, we apply our estimated CASA masks for noise suppression as front-end processing for this comparison system. More specifically, we use the UBMs trained separately in the 3 rooms to perform reverberation classification on noise-suppressed speech, which is consistent with [1] and [29]. The third system, labeled as "Speech Enhancement", uses a state-of-the-art speech enhancement algorithm to suppress noise [14]. We use the source code of this algorithm from the authors to enhance the test speech. The last one, labeled as "Baseline", directly recognizes the test data using MFCC-based anechoic speaker models.

The performance comparison of all these systems along with the proposed system that employs $\mathrm{IBM_R}$ is shown in Fig. 4. The proposed system outperforms all the related systems in all the test conditions. The second best performing system is the reverberation classification method, which is partly due to the effectiveness of the supplied CASA masks. As a state-of-the-art system in noisy and reverberant conditions, the multi-conditional training method does a reasonable job at high SNRs, but not at low SNRs. Although the speech enhancement algorithm was proposed to deal with noisy speech in anechoic conditions, it exhibits reasonable performance in the reverberant environments, as shown by the large improvement over the MFCC baseline, particularly for the engine noise and SSN. It even outperforms multi-conditional training at low SNRs. It could be the smearing effect of late reverberation on speech spectrum is somewhat similar to the corruption by SSN, which can be effectively attenuated by speech enhancement.

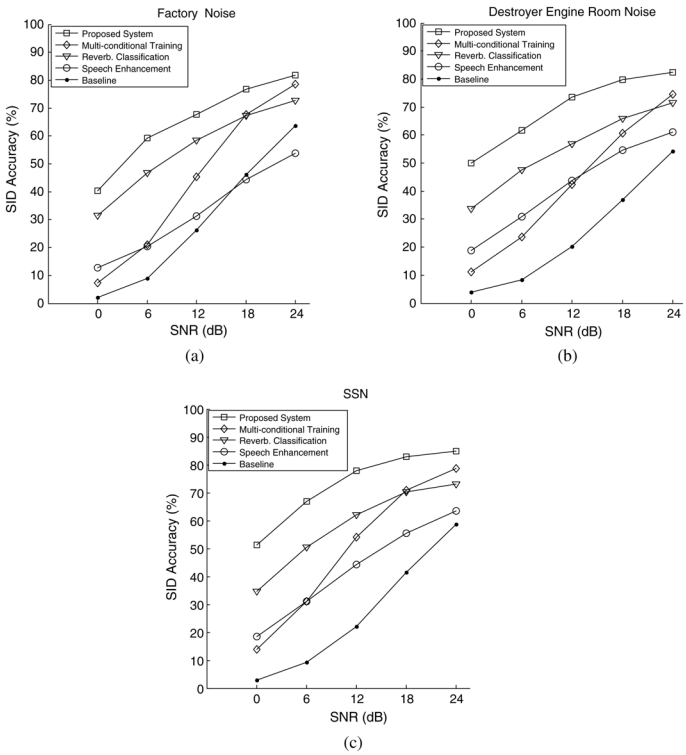

Fig. 4. SID comparisons of the proposed system with related systems under factory noise, destroyer engine room noise and speech shape noise using simulated RIRs.

E. Evaluation with Real Impulse Responses

The results reported so far are generated using simulated RIRs. We now test our system using RIRs recorded in real rooms to assess its utilities in real environments. We use the RIRs collected in Bell Labs [42]. There are three T_{60}'s (300 ms, 500 ms and 900 ms) and 4 RIRs are collected at each T_{60} corresponding to 4 microphone positions. We observe that the actual T_{60} values of these RIRs are probably much higher than the given values according to our measurements. For example, the RIRs of 900 ms would have a T_{60} of $1.4 \sim 1.6$ s as reported in [8]. In this study, we use RIRs from the T_{60}'s of 300 ms and 500 ms. We use the third and fourth RIRs of each T_{60} to create the training set of speaker models. To create the test set, we use the first RIR to convolve with speech and the second one with noise. We then switch these two to get another setting. Therefore we have 2 RIR pairs for each T_{60}. Each test utterance is randomly convolved with 1 of the 2 pairs at each T_{60}. The NIST dataset is employed and the remaining experimental setup stays the same as with simulated RIRs. Note that we use the DNNs trained on simulated RIRs for mask estimation. In other words, there is no retraining of DNNs using real RIRs.

Fig. 5 shows an example of the IBM estimation on real RIRs. As can be seen, the DNNs generalize reasonably well to real RIRs, which is encouraging. This is consistent with [17], where an MLP-based mask estimation algorithm shows similar generalization results. As in Section IV.D., we compare the proposed system with four related systems over 3 noise types and performance is shown in Fig. 6. The proposed system outperforms the

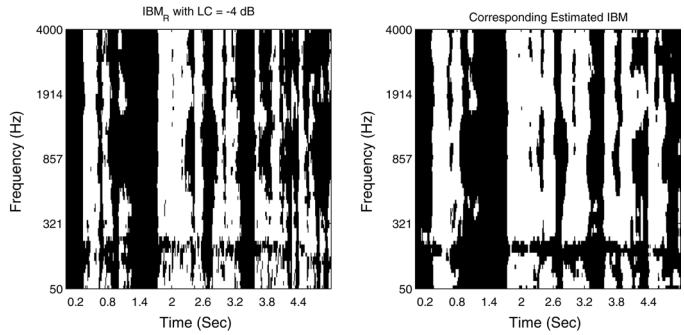

Fig. 5. Comparison of $\mathrm{IBM}_{\mathrm{R}}$ and estimated $\mathrm{IBM}_{\mathrm{R}}$ of an utterance mixed with SSN and real RIR ($\mathrm{T}_{60} = 300$ ms) and $\mathrm{SNR} = 6$ dB.

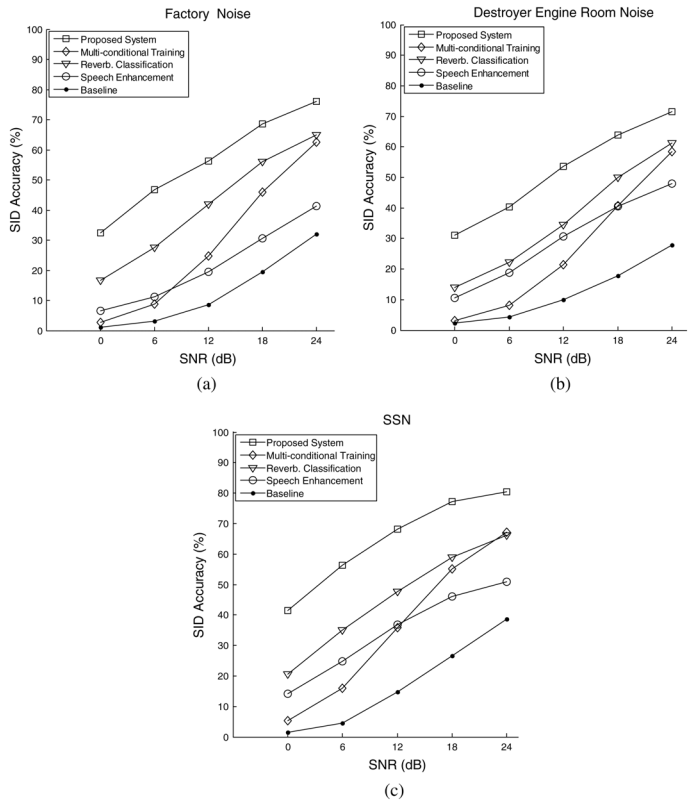

Fig. 6. SID comparisons of the proposed system with related systems under factory noise, destroyer engine room noise and speech shape noise using real RIRs.

related systems in all the test conditions. Compared to simulated RIRs (see Fig. 4), the MFCC baseline is much worse. Even in the noise-free reverberant condition, the MFCC baseline only achieves 56.5% accuracy, which is around 20% lower than with the simulated RIRs. Similarly, the absolute performance of all systems including the proposed system all decreases. This indicates that the real acoustic environments are more challenging than simulated ones for speaker recognition. Overall, the proposed system and the related systems show similar performance trends in simulated and real reverberant environments.

V. Discussion

The combined effects of noise and reverberation have been studied in human listeners [13], [27], and the results indicate that

they pose a greater challenge than individual effects. This study addresses the combined effects in the domain of robust speaker identification. Our benchmark performance suggests that reverberation alone poses a challenge for traditional SID systems already, as shown in the reduced performance of an MFCC baseline from 97.83% to 77.08% using simulated RIRs. Even in the least noisy situation (24 dB SNR), the combined effects further reduce the performance to 58.67% with SSN.

The multi-condition fusion idea investigated here alleviates the problem that it is difficult to accurately match training and testing reverberant conditions. We have observed that the best performing training condition tends to dominate the fusion results, rendering classification of testing reverberant conditions less significant. Module combination further leverages the complementary advantages of noise-robust SID approaches and features, consistent with our previous study on noise-robust SID [43]. The noise susceptibility of MFCC makes it a poor choice for the direct masking module. Nonetheless, its promising results at high SNRs warrant further study to incorporate it into the proposed system.

IBM estimation in noisy and reverberant condition is a very challenging task. Except for [17], little research has been done on this topic. MLPs, SVMs and DNNs have shown promising results on IBM estimation in noisy conditions alone, and this study further considers reverberant conditions by using DNNs due to their performance. The overall mask estimation quality clearly has room to improve, and further improvement can be expected as general IBM estimation progresses.

As shown in Table III, $\mathrm{IBM_{DS}}$ outperforms $\mathrm{IBM_{R}}$. However, estimated IBMs in Table VII outperform estimated $\mathrm{IBM_{DS}}$. $\mathrm{IBM_{R}}$ retains T-F units with significant amounts of late reverberation that are detrimental to recognition due to its noise-like characteristics. On the other hand, $\mathrm{IBM_{DS}}$ is able to capture speech onsets that are relatively robust to reverberation. This may explain the advantage of $\mathrm{IBM_{DS}}$. During mask estimation, the local SNRs of onset-related T-F units are quite low, as indicated by the choice of very low LCs. Therefore the derived features are unlikely discriminative to achieve good mask estimation performance. The T-F units mainly containing late reverberation in $\mathrm{IBM_{R}}$ also do not contain discriminative features. However, the resulting missing errors in $\mathrm{IBM_{R}}$ estimation would not be nearly as harmful as those for $\mathrm{IBM_{DS}}$ estimation. This could explain why estimated $\mathrm{IBM_{R}}$ yields better performance.

We have demonstrated the utilities of our system using RIRs recorded in real environments. It is encouraging to see that DNNs trained on simulated RIRs generalize well to real RIRs. However, we observe that speaker models trained using simulated RIRs perform less well with real RIRs, maybe because speaker models are built from frame-level features which are distorted differently by simulated and real RIRs. On the other hand, mask estimation makes decisions based on energy comparisons, which are not much affected by the differences. Further studies should examine frame-level feature mismatch between simulated and real RIRs.

To conclude, we have investigated the combined effects of noise and reverberation in SID. We employ speaker models trained in multiple reverberant conditions to account for the mismatch created by reverberation. Noise is dealt with using DNN-based CASA separation and two recognition methods which together yield substantial performance improvement over related systems in a wide range of reverberation time and SNRs.

Acknowledgment

We would like to thank Arun Narayanan for his insights on the direct masking part and Ohio Supercomputer Center for their support.

References

[1] A. Akula, V. R. Apsingekar, and P. L. De Leon, "Speaker identification in room reverberation using GMM-UBM," in *Proc. Digital Signal Process. Workshop and 5th IEEE Signal Process. Education Workshop*, 2009, pp. 37–41.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[3] B. Borgstrom and A. McCree, "The linear prediction inverse modulation transfer function (LP-IMTF) filter for spectral enhancement with applications to speaker recognition," in *Proc. ICASSP*, 2012, pp. 4065–4068.

[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] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[6] T. H. Falk and W. Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 90–100, Jan. 2010.

[7] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. ICASSP*, 2012, pp. 4257–4260.

[8] D. Gelbart, "Some resources for noise-robust and channel-robust speech processing," [Online]. Available: http://www.icsi.berkeley.edu/ftp/global/pub/speech/papers/gelbart-ms/pointers/varechoic.zip 2004

[9] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proc. ICSLP*, 1996, pp. 1333–1336.

[10] E. A. P. Habets, "Room impulse response generator," [Online]. Available: http://home.tiscali.nl/ehabets/rir_generator.html 2010

[11] K. Han and D. L. Wang, "A classification based approach to speech segregation," *J. Acoust. Soc. Amer.*, vol. 132, pp. 3475–3483, 2012.

[12] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D. L. Wang, "A direct masking approach to robust ASR," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 1993–2005, Oct. 2013.

[13] O. Hazrati and P. C. Loizou, "The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners," *Int. J. Audiol.*, pp. 437–443, 2012.

[14] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. ICASSP*, 2010, pp. 4266–4269.

[15] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[16] 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.

[17] Z. Jin and D. L. Wang, "Reverberant speech segregation based on multipitch tracking and classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2328–2337, Nov. 2011.

[18] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM-06/08-13, 2005 [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[19] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, pp. 1486–1494, 2009.

[20] P. Krishnamoorthy and S. R. Mahadeva Prasanna, "Application of combined temporal and spectral processing methods for speaker recognition under noisy, reverberant or multi-speaker environment," *Indian Acad. Sci.*, vol. 34, pp. 729–754, 2009.

[21] H. Kuttruff, *Room Acoustics*. New York, NY, USA: Spon, 2000.

[22] Y. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Commun.*, vol. 51, pp. 230–239, 2009.
[23] M. Mandel, S. Bressler, B. Shinn-Cunningham, and D. Ellis, "Evaluating source separation algorithms with reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1872–1883, Sep. 2010.
[24] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2016–2030, 2012 2012.
[25] T. May, S. van de Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 108–121, Jan. 2012.
[26] J. Ming, T. Hazen, J. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.
[27] A. K. Nabelek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.*, vol. 17, pp. 724–739, 1974.
[28] A. Narayanan and D. L. Wang, "The role of binary mask pattern in automatic speech recognition in background noise," *J. Acoust. Soc. Amer.*, vol. 133, pp. 3083–3093, 2013.
[29] I. Peer, B. Rafaely, and Y. Zigel, "Reverberation matching for speaker recognition," in *Proc. ICASSP*, 2008, pp. 4829–4832.
[30] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, 1995.
[31] 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.
[32] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Amer.*, vol. 130, pp. 2153–2161, 2011.
[33] S. Sadjadi and J. Hansen, "Blind reverberation mitigation for robust speaker recognition," in *Proc. ICASSP*, 2012, pp. 4225–4228.
[34] Y. Shao, S. Srinivasan, and D. L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. ICASSP*, 2007, pp. 277–280.
[35] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.
[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, USA: Kluwer, 2005, pp. 181–197.
[37] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: 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] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[40] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.
[41] L. Wang and S. Nakagawa, "Speaker identification/verification for reverberant speech using phase information," in *Proc. WESPAC*, 2009, no. 0130, p. 8.
[42] W. C. Ward, G. W. Elko, R. A. Kubli, and C. McDougald, "The new varechoic chamber at AT&T bell labs," in *Proc. Wallace Clement Sabine Centennial Symp.*, 1994, pp. 343–346.
[43] X. Zhao, Y. Shao, and D. L. Wang, "CASA-based robust speaker identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1608–1616, Jul. 2012.

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.

Y. Wang, photograph and biography not provided at the time of publication.

D. Wang, photograph and biography not provided at the time of publication.