ROBUST SPEAKER IDENTIFICATION IN NOISY AND REVERBERANT CONDITIONS

Xiaojia Zhao[1], *Yuxuan Wang*[1] *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH, USA
{zhaox, wangyuxu, dwang}@cse.ohio-state.edu

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, using 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— Robust speaker identification, noise, reverberation, ideal binary mask, deep neural network

1. 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, computational auditory scene analysis (CASA) was recently employed to remove noise [31]. Speaker features such as modulation spectral features [4] have shown robustness against reverberation. 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 [13] and i-vector based techniques [3].

However, efforts have rarely been made on the combined effects of noise and reverberation. May *et al.* [17] and Gonzalez-Rodriguez *et al*. [6] studied the combined effects using binaural cues and microphone arrays. Garcia-Romero *et al.* [5] and Krishnamoorthy and Prasanna [15] 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 [10], [19].

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 2 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 3, followed by evaluations in Section 4. We conclude this paper in Section 5.

2. SYSTEM OVERVIEW AND FRONT-END PROCESSING

Figure 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) [24]. 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 and subsequently between two modules to make the final SID decision. Below, we describe auditory features and discuss the definition of a CASA mask. Then DNN-based binary masking is described.

2.1. Auditory Features and IBM Definition

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) [31].

Noisy and reverberant speech is first passed through a 64-channel gammatone filterbank to create a two-dimensional *cochleagram* [27]. 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 [31].

A main computational goal of CASA is the *ideal binary mask* (IBM) [26], where each element corresponds to a T-F unit in the cochleagram and indicates whether the local signal-to-noise ratio (SNR) is larger than a threshold called *local criterion* (LC). Given premixed target and interference signals, the IBM can be readily constructed. In this study, the entire reverberant speech is considered as the target and the reverberant noise as interference [12].

This research was supported in part by an AFOSR grant (FA9550-12-1-0130) and an NIDCD grant (R01 DC012048). We would like to thank Ohio Supercomputer Center for their support.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

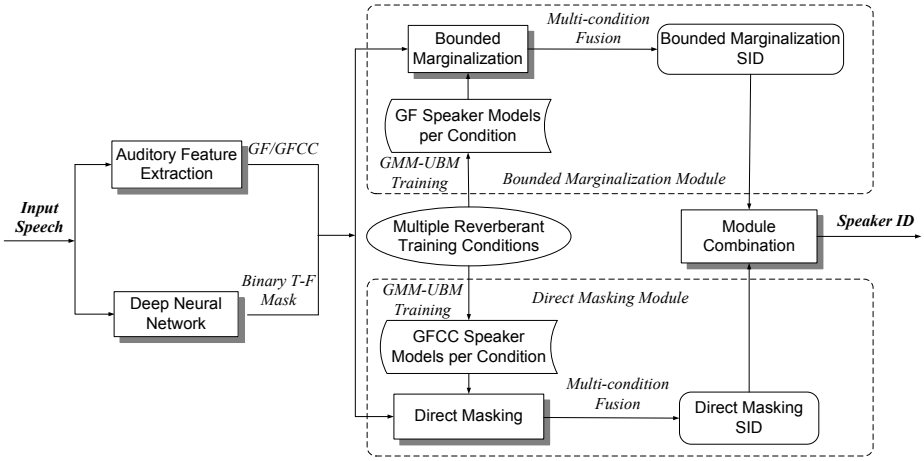

Figure 1. Schematic diagram of the proposed speaker identification system

2.3. 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 [8], [14]. Motivated by their superior performance [28], we employ DNNs for mask estimation in this study.

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. 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:

$$E = \sum_m [d_m \log p_m + (1 - d_m) \log(1 - p_m)] \quad (1)$$

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 [29], which consists of amplitude modulation spectrogram, RASTA-PLP, MFCC and pitch-based features. We used the DNNs described above as the subband classifiers.

3. RECOGNITION METHODOLOGY

The GMM framework along with the universal background model (UBM) [22] is adopted for speaker modeling in this study. 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.

3.1. 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 [18], [31]. Specifically, we perform bounded marginalization on the GF features with a CASA mask specifying the reliable and unreliable T-F units.

3.2. Direct Masking Module

Direct masking is a recently proposed technique for coupling binary masking and speech recognition [9]. 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.

3.3. 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 a room impulse response (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 [16]. 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.

3.4. Multi-condition Fusion and Module Combination

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. 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. We combine the normalized scores using a simple summation. The two modules address SID in noise from different perspectives. GF and GFCC exhibit complementary properties for noise-robust SID [31]. 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 and then simply add the module scores.

4. EVALUATION AND COMPARISON

4.1. Experimental Setup

We randomly drew 300 speakers from the 2008 NIST Speaker Recognition Evaluation dataset (*short2* part of the training set). The telephone conversation excerpt of each speaker is divided 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. We employ the Matlab implementation of the image method to simulate room reverberation [2], [7]; results with recorded impulse responses are given in Section 4.4. The range of T_{60} is varied from 0 to 1s, which covers a broad range of realistic reverberant environments [16]. 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 2m. 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 [22]. 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. 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 2m. 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. Factory noise, speech shape noise (SSN) and destroyer engine room noise from the Noisex-92 database are used as interference [25]. 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 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 [22]. 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 only in selected frames with some target information. Given a frame, it will be selected if its number of reliable units is greater than half number of the channels (i.e. 32) or the median number of reliable T-F units over frames with at least one reliable T-F unit.

For mask estimation, we use two-hidden-layer DNNs, which strike a balance between performance and computational overhead [28]. We train DNNs separately for bounded marginalization and direct masking using noisy and reverberant training data. Used to provide training labels, the IBM is created by comparing 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 [23] and robust speech recognition [20], however, have shown that an LC of 0 dB is not always optimal. To determine the optimal LC for the IBM, we set up another experiment by randomly selecting 50 speakers from the TIMIT corpus. Training and testing data are set up similarly to the NIST dataset. IBMs are derived using different LCs and their corresponding SID performance is compared to determine the optimal LC. The results suggest that an LC of -4 dB is optimal for the bounded marginalization module, and -12 dB for the direct masking module. The corresponding optimal IBMs are employed as target labels for DNN-based mask estimation.

4.2. Performance with Estimated IBM

We now study the utilities of reverberant model training. We first apply anechoic speaker models to the anechoic and reverberant test sets respectively (noise is excluded). As expected, the performance of anechoic speaker models drops substantially in the reverberant test set due to the mismatch. After reverberation is included in speaker models, the performance is shown in Table 1. As shown in the table, the introduction of reverberation in the training data significantly improves performance for all the features.

Features	Anechoic Models	300 ms Models	600 ms Models	900 ms Models
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

Table 1: SID performance (%) of in the reverberant test set. The last three columns represent speaker models trained in the corresponding T_{60} conditions

Now we evaluate the proposed system in the noisy and reverberant test set. 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. As can be seen in Table 2, 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 [31]. 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 modules in every test condition. For example, the combined system outperforms both modules by around 10% for factory noise at 6 dB.

Factory	**0dB**	**6dB**	**12dB**	**18dB**	**24dB**	**Avg.**
MFCC_DM	12.5	23.0	43.5	64.8	75.7	43.9
GFCC_DM	33.3	48.3	61.0	70.3	74.0	57.4
GF_BM	34.1	49.2	59.3	71.9	80.0	58.9
Comb. Syst.	40.3	59.2	67.7	76.8	81.8	65.2
Destroyer	**0dB**	**6dB**	**12dB**	**18dB**	**24dB**	**Avg.**
MFCC_DM	16.2	33.2	50.3	64.2	75.2	47.8
GFCC_DM	35.8	47.3	57.7	66.6	72.1	55.9
GF_BM	45.8	57.9	69.1	78.6	81.8	66.6
Comb. Syst.	50.0	61.7	73.5	79.8	82.4	69.5
SSN	**0dB**	**6dB**	**12dB**	**18dB**	**24dB**	**Avg.**
MFCC_DM	18.5	34.3	55.9	71.3	79.6	51.9
GFCC_DM	37.7	52.9	64.3	70.9	74.9	60.1
GF_BM	44.8	61.2	73.6	79.8	83.4	68.6
Comb. Syst.	51.3	67.0	77.8	83.0	84.8	72.8

Table 2: SID accuracy (%) of the proposed system using estimated IBMs. _DM denotes the direct masking module. _BM denotes the bounded marginalization module. Comb. Syst. denotes the proposed system

4.3. 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 Figure 2, was designed for robust speaker verification using i-vector based techniques [5]. One of the best performing methods in the paper trains Gaussian probabilistic linear discriminant models in both reverberant and noisy conditions. We apply this method to our task. The second system, labeled as "Reverb. Classification", was designed to deal with reverberation alone [1], [21]. 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. 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. The third system, labeled as "Speech Enhancement", uses a state-of-the-art speech enhancement algorithm to suppress noise [11]. 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 is shown in Figure 2. Factory noise is the interference. The proposed system outperforms all the related systems. The second best performing system is the reverberation classification method, which is partly due to the effectiveness of the supplied CASA masks. The multi-conditional training method does a reasonable job at high SNRs, but not at low SNRs. Although the

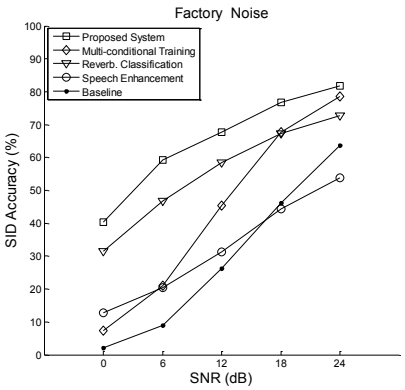

Figure 2. SID comparisons of the proposed system with related systems under factory noise

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 improvement over the MFCC baseline. 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. Evaluations on the other two noise types have shown similar trends.

4.4. 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 real RIRs collected in Bell Labs [30] for training and testing sets. 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.

The results demonstrate that DNNs trained using simulated RIRs generalize reasonably well to real RIRs. The proposed system outperforms the related systems by more than 10% across all the test conditions. This is consistent with simulated RIRs. However, the absolute performance of all systems including the proposed system decreases. This indicates that the real acoustic environments are more challenging than simulated ones for speaker recognition. Detailed results are not presented here due to space limitation.

5. CONCLUDING REMARKS

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 together yield substantial performance improvement over related systems in a wide range of reverberation time and SNRs.

6. RELATION TO PRIOR WORK

The work presented here has focused on the combined effects of noise and reverberation in monaural speaker identification. There has been little study on this problem. The work by Garcia-Romero *et al.* [5] and Krishnamoorthy and Prasanna [15] only reported results in noise and reverberation separately. While the reverberant training idea is related to the Akula et al.'s work [1], we in addition employ CASA to deal with noise.

7. REFERENCES

[1] A. Akula, V.R. Apsingekar and P. L. De Leon, “Speaker identification in room reverberation using GMM-UBM,” Digital *Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, 2009.

[2] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustic Society of America*, vol. 65, pp. 943-950, 1979.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 788 - 798, 2011.

[4] T.H. Falk and W.Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, pp. 90-100, 2010.

[5] 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.

[6] 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.

[7] E.A.P. Habets. (2010) Room Impulse Response Generator. [Online]: home.tiscali.nl/ehabets/rir_generator.html

[8] K. Han and D.L. Wang, “A classification based approach to speech segregation,” *Journal of the Acoustical Society of America*, vol. 132, pp. 3475-3483, 2012.

[9] W. Hartmann, A. Narayanan, E. Fosler-Lussier and D.L. Wang, “A direct masking approach to robust ASR,” *IEEE Tran. Audio, Speech and Language Processing,* vol. 21, pp. 1993-2005, 2013.

[10] O. Hazrati and P. C. Loizou, “The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners,” *Intl J Audiology*, pp. 437-443, 2012.

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

[12] Z. Jin and D.L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, pp. 625-638, 2009.

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

[14] G. Kim, Y. Lu, Y. Hu, and P. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” *Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, 2009.

[15] 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 Academy of Sciences*, vol. 34, pp. 729-754, 2009.

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

[17] 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, and Language Processing*, vol. 20, pp. 2016-2030, 2012.

[18] 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, and Language Processing*, vol. 20, pp. 108-121, 2012.

[19] 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,” *Journal of Speech and Hearing Research*, vol.17, pp. 724-739, 1974.

[20] A. Narayanan and D.L. Wang, “The role of binary mask pattern in automatic speech recognition in background noise,” *Journal of the Acoustical Society of America,* vol. 133, pp. 3083-3093, 2013.

[21] I. Peer, B. Rafaely and Y. Zigel, "Reverberation matching for speaker recognition," in *Proc. ICASSP*, 2008, pp. 4829-4832.

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

[23] N. Roman and J. Woodruff, “Intelligibility of reverberant noisy speech with ideal binary masking,” *Journal of the Acoustical Society of America*, vol. 130, pp. 2153–2161, 2011.

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

[25] 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 Comm.*, vol. 12, pp. 247-251, 1993.

[26] 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 Academic, 2005, pp. 181-197.

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

[28] Y. Wang and D.L. Wang, “Towards scaling up classification-based speech separation,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, pp. 1381-1390, 2013.

[29] Y. Wang, K. Han and D.L. Wang, “Exploring monaural features for classification-based speech segregation”, *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, pp.270-279, 2013.

[30] W.C. Ward, G.W. Elko, R.A. Kubli, and C. McDougald, “The new varechoic chamber at AT&T Bell Labs,” in *Proc. of the Wallace Clement Sabine Centennial Symposium*, 1994, pp. 343-346.

[31] X. Zhao, Y. Shao, and D.L. Wang, “CASA-based robust speaker identification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1608 –1616, 2012.