Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement

Hassan Taherian, Zhong-Qiu Wang, Jorge Chang, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Deep neural network (DNN) embeddings for speaker recognition have recently attracted much attention. Compared to i-vectors, they are more robust to noise and room reverberation as DNNs leverage large-scale training. This article addresses the question of whether speech enhancement approaches are still useful when DNN embeddings are used for speaker recognition. We investigate single- and multi-channel speech enhancement for text-independent speaker verification based on x-vectors in conditions where strong diffuse noise and reverberation are both present. Single-channel (monaural) speech enhancement is based on complex spectral mapping and is applied to individual microphones. We use masking-based minimum variance distortion-less response (MVDR) beamformer and its rank-1 approximation for multi-channel speech enhancement. We propose a novel method of deriving time-frequency masks from the estimated complex spectrogram. In addition, we investigate gammatone frequency cepstral coefficients (GFCCs) as robust speaker features. Systematic evaluations and comparisons on the NIST SRE 2010 retransmitted corpus show that both monaural and multi-channel speech enhancement significantly outperform x-vector's performance, and our covariance matrix estimate is effective for the MVDR beamformer.**

***Index Terms*—Robust speaker recognition, speech enhancement, masking-based beamforming, x-vector, gammatone frequency cepstral coefficient (GFCC).**

I. Introduction

THE RAPID deployment of smart speaker devices such as Amazon Echo and Google Home has propelled the utilization of speaker recognition (SR) systems. It is useful for these devices to authenticate the claimed identity of a user based on previous enrollment speech data, a task known as speaker verification. SR can be either text-dependent, where the speech content is known a priori, or text-independent. Robust SR is of critical importance as background noise and room reverberation can severely degrade the performance of such systems [12], [17]. In this paper, we address the robustness in text-independent speaker verification.

The main approach to robust SR is based on the i-vector framework [5]. This approach includes a front-end processing where acoustic features are projected into a low dimensional space and a probabilistic linear discriminant analysis (PLDA) [16] classifier as the back-end. Notable improvements on the front-end include the replacement of the Gaussian mixture model (GMM) component of i-vectors by DNNs [18] and the combination of acoustic features and bottleneck features extracted from DNN [20], [21]. The back-end of SR also has been improved by introduction of a signal-to-noise ratio (SNR) invariant version of PLDA [19] and multi-condition training [17].

Recently, several studies have employed an end-to-end DNN based SR system instead of i-vectors to directly discriminate speakers and demonstrated their potential in different tasks [13], [30]. Following this approach, Snyder *et al.* proposed the x-vector representation, a fixed-dimensional DNN embedding, to substitute the i-vector front-end [37]. Aside from reducing the complexity of end-to-end DNNs and utilization of the algorithms associated with the back-end like PLDA and length normalization techniques, the x-vector approach significantly improves the robustness against noise and reverberation by leveraging data augmentation [25], [37]. With noise variability reduced by DNN embeddings, is speech enhancement needed for further increasing the robustness?

Speech enhancement has been used to increase the robustness of SR systems. Depending on the number of available microphones, speech enhancement can be either monaural or multi-channel. As SR has traditionally been applied to telephone speech, monaural speech enhancement has been mostly investigated. In [2], a DNN is trained for the ideal ratio mask (IRM) estimation to attenuate background noise and subsequently enhanced speech is fed to an i-vector based SR system. The use of deep autoencoders to perform speech enhancement for i-vectors is presented in [27] and is later extended to x-vectors by Novotnỳ *et al.* [24]. Speech enhancement based on conditional generative adversarial networks for speaker verification is investigated in [22]. In [34], a loss function for speech enhancement was proposed on the basis of speaker verification feedback. Joint optimization of speech enhancement and end-to-end DNN based

Manuscript received November 6, 2019; revised February 4, 2020 and March 11, 2020; accepted April 5, 2020. Date of publication April 13, 2020; date of current version May 5, 2020. This work was supported in part by the National Science Foundation under Grant ECCS-1808932 and in part by Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zheng-Hua Tan PhD. *(Corresponding author: Hassan Taherian.)*

Hassan Taherian and Zhong-Qiu Wang are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: taherian.1@osu.edu; wangzhon@cse.ohio-state.edu).

Jorge Chang is with the Department of Psychology, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: changcheng.1@osu.edu).

DeLiang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210-1277 USA, and also with the Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2020.2986896

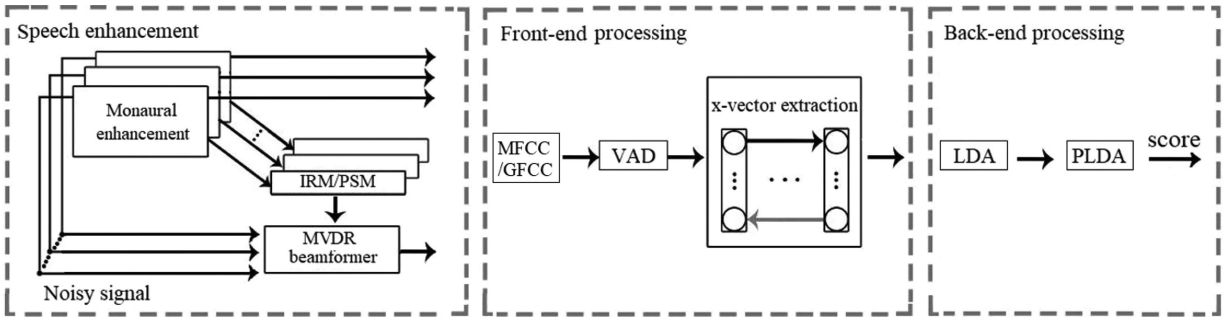

Fig. 1. Schematic diagram of the proposed robust speaker recognition framework.

SR [52] is another recent contributions that shows the usefulness of speech enhancement.

Smart speakers use multiple microphones. By exploiting spatial information afforded by a microphone array, one can obtain a greater speech ehancement improvement in terms of speech intelligibility and quality compared to using a single microphone [10], [42]. In recent years multi-channel speech enhancement has experienced major advances by incorporating DNN based time-frequency (T-F) masking into adaptive beamforming, a widely utilized technique in this domain [14], [49], [51]. Adaptive beamformers are based on second-order statistics of speech and noise, and the more accurate estimation of these statistics by DNN is the main reason behind the recent progress. Mošner *et al.* recently studied the effects of multi-channel speech enhancement on far-field speaker recognition in reverberant environments [23]. In their study, a combination of multi-channel weighted prediction error (WPE) [50] and masking-based beamforming is used to enhance the quality of reverberant speech signal before extracting features for the SR system. Their analysis, however, is based on i-vectors, and the effectiveness of multi-channel speech enhancement is unclear for x-vector based SR.

In a preliminary study, we recently examined different masking-based beamforming methods in conditions where target speech is corrupted by strong diffuse noise and room reverberation and evaluated their performance on the i-vector and x-vector based SR [39]. In the present study, we further develop our approach by providing in-depth comparisons between monaural and multi-channel speech enhancement and their application to robust SR. Moreover, we propose a new method to estimate the speech and noise statistics for masking-based beamforming on the basis of the recently introduced gated convolutional recurrent network (GCRN) [40]. A diagram of the proposed framework is shown in Fig. 1. For monaural speech enhancement, we employ GCRN to perform complex spectral mapping to predict the real and imaginary spectra of clean speech. We investigate the enhancement capabilities of GCRN in different SNR conditions for i-vector and x-vector based SR. Then the performance of multi-channel speech enhancement is examined for i-vectors and x-vectors, where we estimate the IRM and the phase sensitive mask (PSM) from estimated spectra to compute the speech and noise covariance matrices. The computed statistics are then used to calculate steering vectors for MVDR and rank-1 approximated MVDR beamformers.

We also incorporate gammatone frequency cepstral coefficients (GFCCs) as speaker features for training i-vectors and x-vectors and demonstrate their robustness over commonly used mel-frequency cepstral coefficients (MFCCs) under noisy and reverberant conditions.

The rest of the paper is organized as follows. In Section II, we describe monaural speech enhancement and GCRN architecture. Section III proposes a new masking-based beamformer based on MVDR and rank-1 approximated MVDR. Section IV explains GFCC feature extraction for i-vectors and x-vectors. We then provide the experimental setup and the evaluation results in Section V and VI, respectively. Concluding remarks are given in Section VII.

II. Monaural Speech Enhancement

A noisy speech signal received by an individual microphone can be expressed with the short-time Fourier transform (STFT):

$$y(t,f) = a(f)s(t,f) + n(t,f) \tag{1}$$

where $s(t,f)$ is the STFT representation of speech signal at frame t and frequency f, $a(f)$ is the time-invariant acoustic transfer function (ATF), and $y(t,f)$ and $n(t,f)$ represent the STFT of noisy mixture and noise, respectively. In the case of reverberation, we can decompose the ATF into two parts and write Eq. (1) as:

$$y(t,f) = c(f)s(t,f) + h(t,f) + n(t,f) \tag{2}$$

where $c(f)$ is the ATF from the speech source to the microphone, $c(f)s(t,f)$ represents the direct-path speech signal and $h(t,f)$ its reverberation. Monaural speech enhancement formulates the estimation of clean speech $c(f)s(t,f)$ from noisy mixture $y(t,f)$ as a supervised learning problem [41]. In supervised speech enhancement, different training targets have been proposed [42], [43], inspired by the T-F masking concept from computational auditory scene analysis (CASA) [41]. As a widely used training target, the IRM is defined as the ratio of the energy of clean and noisy speech at each T-F unit:

$$\mathrm{IRM}(t,f) = \frac{|c(f)s(t,f)|^2}{|c(f)s(t,f)|^2 + |h(t,f) + n(t,f)|^2}. \tag{3}$$

Once the IRM is learned, clean speech can be estimated by applying the estimated IRM to the STFT of the noisy mixture. It

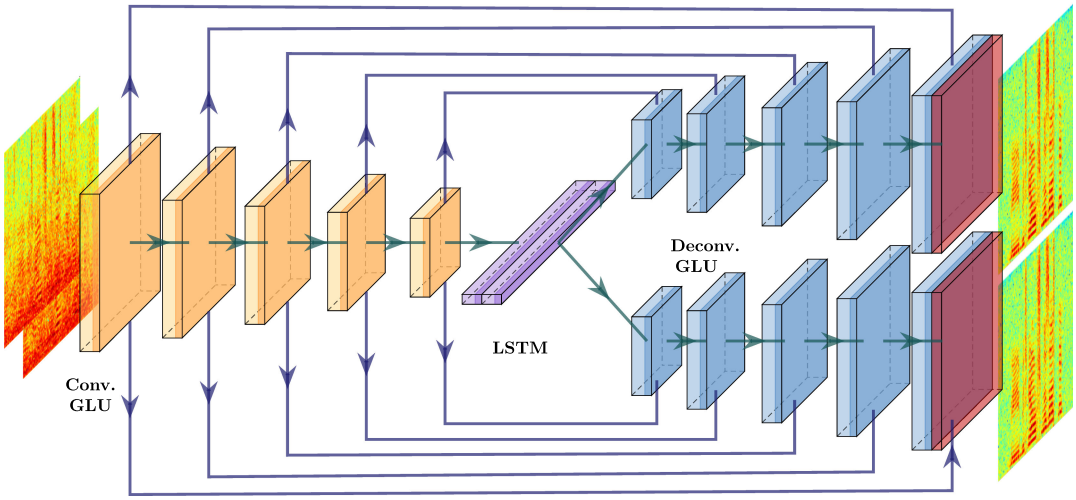

Fig. 2. Network architecture of GCRN for complex spectral mapping. Skip connections link encoders to the corresponding decoders. 'Conv' and 'Deconv' denote convolution and deconvolution, respectively.

is worth noting that the IRM only enhances the magnitude spectrogram of the noisy mixture and leaves its phase unchanged. While phase is important for speech quality [26], the lack of structure in the phase spectrogram makes its direct estimation intractable [47]. Recently, however, new training targets have been proposed to utilize the spectrotemporal structure of real and imaginary components of a speech signal in order to enhance both magnitude and phase [9], [47]. In [40], Tan *et al.* proposed GCRN to map real and imaginary components of a noisy mixture to the corresponding components of clean speech.

GCRN is based on an encoder-decoder architecture that includes a recurrent neural network with long short-term memory (LSTM) between the encoder and the decoder. Fig. 2 illustrates GCRN network architecture [40]. The encoder comprises five convolutional gated linear unit (GLU) blocks [4] and two LSTM layers, followed by two distinct decoder modules each of which has five deconvolutional GLU blocks and one linear layer to predict the real and imaginary spectra of clean speech. The input contains two parts representing real and imaginary components of the noisy mixture. The number of kernels is progressively doubled in the encoder and halved in the decoder to guarantee that the output has the same T-F representation as the input. GLU blocks are accompanied by a batch normalization layer and an exponential linear unit activation function. The trainable parameters of the LSTM layers are reduced by using a group strategy and a parameter-free representation rearrangement layer.

III. Multi-Channel Speech Enhancement

A. Rank-1 Approximated MVDR Beamformer

In this section, we extend the signal model introduced in Section II to multiple microphones. Assuming there are M microphones forming a microphone array, Eq. (2) can be extended to:

$$\boldsymbol{y}(t,f) = \boldsymbol{c}(f)s(t,f) + \boldsymbol{h}(t,f) + \boldsymbol{n}(t,f) \tag{4}$$

where $\boldsymbol{c}(f) \in \mathbb{C}^M$ denotes a steering vector that contains information about inter-microphone level and phase differences of direct-path speech signal [10], $\boldsymbol{y}(t,f)$, $\boldsymbol{h}(t,f)$ and $\boldsymbol{n}(t,f) \in \mathbb{C}^M$ respectively represent the STFT vectors of the noisy mixture, speech reverberation and noise.

Beamforming applies a frequency-dependent linear spatial filter $\boldsymbol{w}(f) \in \mathbb{C}^M$ to the noisy mixture to suppress noise and reverberation. Adaptive beamformers such as MVDR are based on input signal statistics, including the covariance matrix of the noisy mixture defined as:

$$\begin{aligned} \boldsymbol{\Phi}_y(f) &= \mathbb{E}\left\{\boldsymbol{y}(t,f)\boldsymbol{y}(t,f)^H\right\} \\ &= \boldsymbol{\Phi}_s(f) + \underbrace{\boldsymbol{\Phi}_h(f) + \boldsymbol{\Phi}_n(f)}_{\boldsymbol{\Phi}_N(f)} \end{aligned} \tag{5}$$

where $\mathbb{E}\{.\}$ is the expectation operator, T is the total number of frames used in the summation, and $(.)^H$ denotes conjugate transpose. $\boldsymbol{\Phi}_s(f)$ and $\boldsymbol{\Phi}_N(f)$ are the covariance matrices of clean speech and overall interference, respectively. The above equation assumes that the components of Eq. (4) are uncorrelated.

In the MVDR beamformer, the optimal $\boldsymbol{w}(f)$ is found by minimizing the variance of the beamformer's output without distorting the speech signal. Mathematically, $\boldsymbol{w}(f)$ is realized by solving the constrained optimization below [8]:

$$\begin{aligned} &\underset{\boldsymbol{w}(f)}{\text{argmin}} \;\; \boldsymbol{w}(f)^H \boldsymbol{\Phi}_N(f) \boldsymbol{w}(f) \\ &\text{subject to} \quad \boldsymbol{w}(f)^H \boldsymbol{c}(f) = 1 \end{aligned} \tag{6}$$

and the closed-form solution can be expressed as:

$$\boldsymbol{w_{opt}}(f) = \frac{\boldsymbol{\Phi}_N^{-1}(f)\boldsymbol{c}(f)}{\boldsymbol{c}(f)^H \boldsymbol{\Phi}_N^{-1}(f)\boldsymbol{c}(f)}. \tag{7}$$

As Eq. (7) suggests, the accurate estimation of the steering vector and the covariance matrix of interference is key to MVDR beamforming. Recent studies employ deep learning based T-F masking for estimation of the covariance matrices and report substantial improvements in beamforming performance [14], [15], [45]. In this approach, a T-F mask is used to weigh the

mixture covariance matrix at each T-F unit, and hence the speech or noise statistics are collected from speech-dominant or noise-dominant T-F units [42]. Concretely, the covariance matrices of speech and interference for the MVDR beamformer are estimated as [42]:

$$\mathbf{\Phi}_N(f) = \frac{\sum_t (1 - m(t,f)) \boldsymbol{y}(t,f) \boldsymbol{y}(t,f)^H}{\sum_t (1 - m(t,f))} \tag{8}$$

$$\mathbf{\Phi}_y(f) = \frac{1}{T} \sum_t \boldsymbol{y}(t,f) \boldsymbol{y}(t,f)^H \tag{9}$$

$$\mathbf{\Phi}_s(f) = \mathbf{\Phi}_y(f) - \mathbf{\Phi}_N(f) \tag{10}$$

where $m(t,f)$ represents the T-F mask. A single mask is estimated from each microphone independently, then $m(t,f)$ is derived by combining all T-F masks using median pooling [14].

The principal eigenvector of the estimated speech covariance matrix is a valid estimate of the steering vector, as $\mathbf{\Phi}_s(f)$ is a rank-1 matrix by definition. However, in practice, the estimation of $\mathbf{\Phi}_s(f)$ usually does not result in a rank-1 matrix, especially in the presence of reverberation. To prevent the erroneous estimation of the steering vector, we can obtain a rank-1 matrix from the speech covariance matrix using low rank approximation [32], [38], [46]:

$$\mathbf{\Phi}_s(f) = \mathbf{\Phi}_{r1}(f) + \mathbf{\Phi}_z(f) \tag{11}$$

where $\mathbf{\Phi}_{r1}(f)$ is a rank-1 matrix and $\mathbf{\Phi}_z(f)$ is the remainder matrix. Serizel *et al.* [32] proposed several techniques for estimating $\mathbf{\Phi}_{r1}(f)$, namely, first column decomposition, eigenvalue decomposition (EVD) and generalized eigenvalue decomposition (GEVD). In GEVD, we jointly diagonalize $\mathbf{\Phi}_s(f)$ and $\mathbf{\Phi}_N(f)$:

$$\begin{cases} \mathbf{\Phi}_s(f) = \boldsymbol{Q}(f) \Lambda(f) \boldsymbol{Q}(f)^H \\ \mathbf{\Phi}_N(f) = \boldsymbol{Q}(f) \boldsymbol{I}_M \boldsymbol{Q}(f)^H \end{cases} \tag{12}$$

where $\boldsymbol{I}_M$ is an $M \times M$ identity matrix and $\Lambda(f) = \mathrm{diag}\{\lambda_1(f), \ldots, \lambda_M(f)\}$, assuming that eigenvalues are sorted in the descending order. Then, the covariance matrices in Eq. (11) can be written as:

$$\mathbf{\Phi}_s(f) = \boldsymbol{Q}(f) \mathrm{diag}\{\lambda_1(f), \ldots, \lambda_M(f)\} \boldsymbol{Q}(f)^H \tag{13}$$

$$\mathbf{\Phi}_{r1}(f) = \boldsymbol{Q}(f) \mathrm{diag}\{\lambda_1(f), 0, \ldots, 0\} \boldsymbol{Q}(f)^H \tag{14}$$

$$\mathbf{\Phi}_z(f) = \boldsymbol{Q}(f) \mathrm{diag}\{0, \lambda_2(f), \ldots, \lambda_M(f)\} \boldsymbol{Q}(f)^H \tag{15}$$

$\mathbf{\Phi}_z(f)$ can be interpreted as residual noise error and ignored. Hence, $\mathbf{\Phi}_s(f)$ simplifies to [46]:

$$\mathbf{\Phi}_{r1}(f) = \frac{tr(\mathbf{\Phi}_s(f))}{tr(\boldsymbol{q}_1(f) \boldsymbol{q}_1(f)^H)} \boldsymbol{q}_1(f) \boldsymbol{q}_1(f)^H \tag{16}$$

where $\boldsymbol{q}_1(f)$ is the first column of $\boldsymbol{Q}(f)$ and $tr(.)$ is the trace operator. The steering vector $\boldsymbol{c}(f)$ can be more accurately estimated as the principal eigenvector of $\mathbf{\Phi}_{r1}(f)$ since the rank-1 assumption is guaranteed to be valid.

B. T-F Masking

The performance of masking-based beamformers depends on accurate T-F mask estimation. The mask definition impacts the estimation of the steering vector and the interference covariance matrix. In [15], the T-F mask is regarded as the probability obtained from a complex Gaussian mixture model. Other studies employ DNN to estimate the ideal binary mask [14], the IRM [51] and the complex IRM [48] for the estimation of the speech and interference covariance matrices.

As GCRN estimates both real and imaginary spectra of clean speech we can construct different T-F masks without modifying the network training target. In the case of the IRM, we define:

$$\mathrm{IRM}_i(t,f) = \frac{|\hat{s}_i(t,f)|^2}{|\hat{s}_i(t,f)|^2 + |\hat{n}_i(t,f)|^2} \tag{17}$$

where $\hat{s}_i(t,f)$ is GCRN's estimated spectrogram of clean speech obtained from microphone i and

$$\hat{n}_i(t,f) = y_i(t,f) - \hat{s}_i(t,f) \tag{18}$$

is the estimated interference. The IRM defined in Eq. (17) only uses the magnitude information of GCRN's estimated spectrogram. In [40] it is shown that GCRN can provide a phase estimate of clean speech as well. Therefore we also use the PSM by incorporating the phase information for estimating the T-F mask as [6]:

$$\mathrm{PSM}_i(t,f) = Re\left(\frac{\hat{s}_i(t,f)}{y_i(t,f)}\right) = \frac{|\hat{s}_i(t,f)|}{|y_i(t,f)|} \cos(\theta_i) \tag{19}$$

where θ_i is the phase difference between the noisy mixture and GCRN's estimated spectrogram of clean speech. To simplify PSM estimation, we truncate it to have a value between 0 and 1.

IV. Robust Speaker Features

MFCC is the most commonly used feature for speaker recognition. Shao *et al.*, however, reported that GFCC performs better than MFCC for speaker identification when input signal is corrupted by background noise [31], [33], [53], [54]. GFCC is based on the gammatone filterbank which is derived from psychophysical observations of human cochlear filtering [41].

We investigate GFCC as an alternative speaker feature, on the basis of a 64-channel gammatone filterbank whose center frequencies range from 20 to 3700 Hz. Each filter response is fully rectified and downsampled to 100 Hz along the time dimension, corresponding to a frame rate of 10 ms. The magnitude of the downsampled signals is then loudness-compressed by a cubic root operation. The resulting gammatone feature (GF) matrix is a variant of the cochleagram [41], analogous to the widely used spectrogram. Fig. 3 contrasts the representations of the cochleagram and spectrogram of a clean speech utterance. As shown in the figure, unlike the linear frequency resolution of the spectrogram, there are more filters and finer frequency resolutions at low frequencies compared to high frequencies on the cochleagram.

The cochleagram responses of neighboring filters are correlated because of their bandwidth overlap. This correlation is removed by applying discrete cosine transform to the GF matrix. The extracted features are called gammatone frequency cepstral coefficients [33], [53]. Rigorously speaking, GFCCs

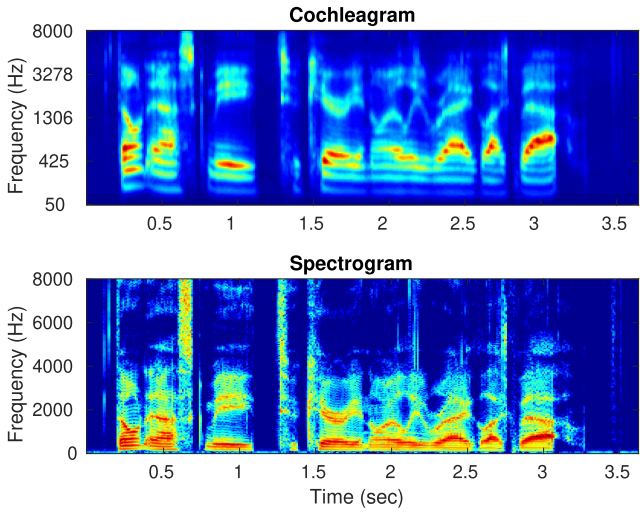

Fig. 3. The comparison of cochleagram and spectrogram for a clean speech utterance. The frequency resolution at low and high frequencies in the cochleagram is different, unlike the spectrogram.

are not cepstral coefficients as cepstral analysis requires a log operation between the first and the second frequency analysis for the deconvolution purpose [33]. The term cepstral coefficients in GFCC merely refers to functional similarities in GFCC and MFCC derivations.

V. Experimental Setup

A. Dataset

We evaluate our monaural and multi-channel systems on real and simulated room impulse responses (RIRs). The real RIR experiments are based on the recently-proposed NIST retransmitted dataset [23], which includes 459 recordings uttered by totally 150 female speakers with having either three or five minute duration for each recording. In this dataset, a loud speaker is employed to retransmit the utterances in a highly reverberant environment.[1] Out of 14 available microphones, 6 are placed for beamforming purposes. Their placement forms an ad-hoc microphone array with large inter-microphone distance in the range of 2.80 to 7.62 meters. The microphone placement is depicted in Fig. 4 a. Note that the plane wave assumption does not hold in this setup since the aperture of the array is the same order of magnitude as the distance from the loud speaker to microphones. This implies that conventional beamformers such as delay-and-sum which are based on estimation of direction of arrival, may not perform optimally. This hypothesis is confirmed in [23] by showing that conventional beamformers do not perform better than a single microphone in speaker recognition task. In contrast, the masking-based beamforming is shown to be beneficial since it does not consider any prior knowledge such as the array geometry or the plane wave assumption [15].

For the simulated RIRs experiments, we use the anechoic version of the NIST retransmitted dataset and convolve it with RIRs

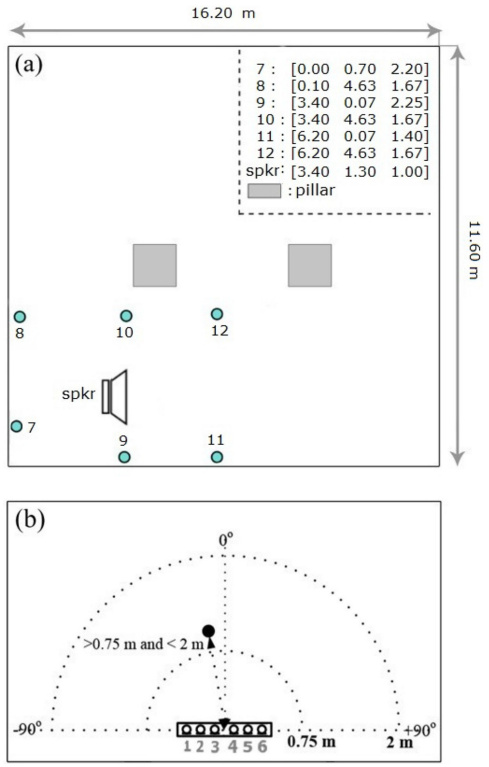

Fig. 4. Illustration of the experimental setup with (a) real RIRs (from [23]), and (b) simulated RIRs. The distance from a speaker to the array center is between [0.75,2] m for simulated RIRs.

generated by the image method.[2] We follow the setup described in [44] for sampling the parameters of the RIR simulation. An illustration of this setup is depicted in Fig. 4 b corresponding to a uniform linear microphone array with 6 microphones and randomly generated rooms and speaker locations. The reverberation time (T60) and inter-microphone distance are selected randomly in the range of 0.4 to 0.8 seconds and 0.02 to 0.09 meters, respectively.

We mix the real and simulated RIR datasets with diffuse babble noise to emulate a challenging background noise without directional information. To create the babble noise, 10 utterances are randomly selected from the TIMIT dataset, and mixed together. We repeat this procedure for all utterances, and then concatenate all mixtures to generate 80-minute babble noise. The babble noise is split into two halves for training and testing. To make the babble noise diffuse, we use the algorithm in [11] that generates sensor signals based on a predefined spatial coherence constraint induced by the array geometry. The sampling rate for all utterances is 8 kHz.

B. Speech Enhancement

We train an LSTM based recurrent neural network for IRM estimation as a baseline to compare with GCRN. The frame length and frame shift for both networks are set to 40 ms and 20 ms, respectively. The input feature for LSTM is 161 log

[1]The authors did not report reverberation time (T60), but it can be inferred by listening.

[2][Online]. Available: https://github.com/ehabets/RIR-Generator

magnitude spectrogram after global mean-variance normalization. The LSTM network includes 4 hidden layers each with 600 units, and an output layer with 161 sigmoidal units. Mean squared error is used as the cost function for both LSTM and GCRN.

The AMSGrad optimizer [29] is used for training GCRN with a learning rate of 0.0001. The two networks have comparable numbers of training parameters with 9.77 and 10.50 million for GCRN and LSTM, respectively. Both DNNs are trained on a subset of NIST SRE 2008 [1] which includes three- or five-minute telephone or interview conversations from female speakers. The total duration of the training data is 140 hours. The RIRs are convolved with the training dataset with T60 between 0.2 and 1 seconds using the simulation procedure described in Section V-A. Then, we mix the training data with the diffuse babble noise at SNRs ranging from 0 dB to 15 dB with reverberant speech treated as the reference signal in SNR calculation.

C. Speaker Recognition

We conduct our SR experiments using i-vectors [36] and x-vectors [37], both are implemented in Kaldi [28]. To train i-vectors and x-vectors, we select 86,629 utterances from the PRISM dataset [7]. We augment the training data by adding two replicas from reverberation and babble noise. The reverberation replica is generated by convolving small and medium room RIRs from the OpenSLR dataset with clean training utterances. The babble noise replica is generated by randomly selecting 3–7 utterances from the MUSAN speech dataset [35] and then mixed with clean speech training data with SNRs in the range of 6–13 dB. It should be noted while the data augmentation significantly improves x-vector performance [37], it is shown that i-vectors might not benefit from data augmentation since i-vectors are trained in an unsupervised fashion and adding noisy data forces GMM to model noise variability instead of speaker variability [24]. Therefore, only clean data is used for training i-vectors. For x-vectors, we select 128,000 utterances randomly from both replicas and add to the clean training dataset. Furthermore, we remove those utterances that are shorter than 500 frames and those speakers that have fewer than 8 utterances. Totally, 169,660 utterances are generated for x-vector training and they include 5,565 speakers and 10,381 hours of data.

I-vectors based on MFCC use 23 mel-filters with triangular overlapping windows in the range of 20 to 3700 Hz. Twenty-dimensional MFCC features are calculated every 10 ms with a window length of 20 ms. For i-vectors based on GFCC, we follow the procedure described in Section IV for extracting 20-dimensional GFCC features. After feature extraction, delta and acceleration features are added to create a 60-dimensional feature. Cepstral mean normalization (CMN) is applied for a sliding window of 3 s. The default energy-based voice activity detector (VAD) in Kaldi is applied. The i-vector extractor is trained with 2048-component universal background model (UBM) based on a full-covariance GMM. We use a portion of the training data (15,600 utterances) for UBM training. Linear discriminant analysis (LDA) is applied to i-vectors to reduce the dimensionality from 600 to 200. Prior to PLDA scoring, i-vectors are centered around a global mean followed by length normalization.

TABLE I
MONAURAL SPEAKER VERIFICATION RESULTS (%EER) WITH MFCC AS THE INPUT FEATURE. RESULTS ARE OBTAINED BY AVERAGING OVER ALL MICROPHONES. 'SIMU' REFERS TO SIMULATED RIRS

		i-vector			x-vector		
RIR	SNR	Noisy	LSTM	GCRN	Noisy	LSTM	GCRN
Simu	0 dB	19.43	12.66	11.01	9.92	8.12	7.30
	5 dB	10.97	7.95	6.53	4.89	4.82	4.65
	10 dB	6.85	5.44	4.54	3.44	3.60	3.65
	15 dB	5.23	4.39	3.95	2.95	3.21	3.30
Real	0 dB	37.61	31.61	28.13	28.37	25.94	24.51
	5 dB	29.44	24.72	21.26	21.00	18.80	17.72
	10 dB	22.08	18.36	16.21	15.23	13.42	12.96
	15 dB	16.06	13.91	13.89	11.44	9.96	10.62

For x-vectors based on MFCC, input features are 23-dimensional MFCC, with the 25-ms frame length and a 10-ms frame shift. To have a fair comparison, we use the lowest 23-order GFCC features for training x-vectors. Subsequently, the same VAD and CMN with 3 s sliding window are applied. Similar to the original Kaldi recipe [37], 512-dimensional embeddings are extracted from the affine component of the first segment-level layer of x-vector's DNN. The dimensionality of x-vectors is reduced to 150 by LDA and they are subject to length normalization before PLDA scoring.

VI. EXPERIMENTAL RESULTS AND COMPARISONS

A. Speaker Verification Based on Single-Channel Speech Enhancement

We report speaker verification results in terms of equal error-rate (EER). The i-vector and x-vector system based on MFCC score 2.5% and 1.88% EER for clean test data (anechoic and noise free), respectively. With GFCC as the input feature, we achieve 2.80% and 1.67% EER for i-vectors and x-vectors respectively. Table I displays the average EER score over 6 microphones with simulated and real RIRs when MFCC is used. The unprocessed utterances and the enhanced utterances by LSTM or GCRN are reported in 4 different SNR conditions. It can be observed that the i-vector method benefits from speech enhancement in all scenarios. The benefit is the largest at the least favorable SNR, where EER scores for the unprocessed utterances are the highest. At the SNR of 0 dB, mean unprocessed scores average to 19.43% in the simulated RIRs. This value is decreased to 12.66% and 11.01% by using LSTM and GCRN, respectively. We can see the same trend as the SNR increases for both simulated and real RIRs. At higher SNRs, speech enhancement still improves the i-vector performance, although to smaller extents. With regard to the comparison between the two speech enhancement systems, GCRN outperforms LSTM in all conditions except for the real RIR at 15 dB SNR where LSTM and GCRN perform comparably.

For x-vectors, we observe that they yield significantly better EER scores in all conditions compared to i-vectors. Note that,

TABLE II
MONAURAL SPEAKER VERIFICATION RESULTS (%EER) WITH GFCC AS THE INPUT FEATURE. RESULTS ARE OBTAINED BY AVERAGING OVER ALL MICROPHONES

RIR	SNR	i-vector			x-vector		
		Noisy	LSTM	GCRN	Noisy	LSTM	GCRN
Simu	0 dB	13.84	10.04	9.38	7.48	6.78	7.02
	5 dB	8.42	6.81	6.43	4.18	4.42	4.70
	10 dB	6.10	5.59	5.42	3.11	3.51	3.63
	15 dB	5.17	5.03	4.89	2.88	3.14	3.20
Real	0 dB	27.45	24.51	21.75	24.23	21.58	20.51
	5 dB	21.44	19.39	16.41	16.96	15.94	14.57
	10 dB	16.84	15.81	13.38	12.09	11.69	11.02
	15 dB	13.85	13.26	12.11	9.24	8.96	9.22

TABLE III
MONAURAL SPEAKER VERIFICATION RESULTS (%EER) USING WPE PREPROCESSING WITH REAL RIRS. RESULTS ARE OBTAINED BY AVERAGING OVER ALL MICROPHONES. 'SE' REFERS TO SPEECH EHANCEMENT WITH GCRN

	MFCC				GFCC			
	i-vector		x-vector		i-vector		x-vector	
SNR	WPE	WPE +SE	WPE	WPE +SE	WPE	WPE +SE	WPE	WPE +SE
0 dB	37.04	25.70	27.25	21.37	26.92	19.72	23.01	18.90
5 dB	29.04	19.78	19.57	15.45	20.16	15.31	15.71	13.61
10 dB	20.81	15.43	13.52	11.83	15.43	12.51	11.11	10.26
15 dB	14.68	12.98	9.99	9.78	12.30	10.85	8.18	8.65

although the SNR range for x-vector training is 6–13 dB, the x-vector system shows robustness in lower SNR conditions. Another comparison of interest involves the performance of x-vectors after speech enhancement. As shown in Table I, enhancing noisy speech signal before extracting x-vectors is effective, especially in low SNR conditions. However, unlike i-vectors, speech enhancement causes a little degradation at higher SNRs in simulated RIRs where reverberation is not as challenging as real RIRs. A possible solution would be to apply speech enhancement during x-vector training which can result in EER reduction as reported in [24]. On the other hand, in the real RIR conditions, speech enhancement improves EER results at all SNR levels.

One could argue that increasing the amount of training data by including more SNR levels may lessen the need for speech enhancement, as x-vectors benefit from data augmentation. However, generating noisy replicas in the x-vector requires each speaker to be mixed with a number of SNR conditions in order to exhibit SNR generalization. This can lead to a huge amount of training data and be very time consuming for training. On the other hand, generalization to unseen conditions is possible for speech enhancement with a relatively small amount of training data compared to the x-vector training [3], [40].

We present the performance of monaural speaker verification based on GFCC features in Table II. The pattern of the results in Table II is similar to that in Table I. However, GFCC features show consistent improvements over MFCC features for both noisy and enhanced utterances, consistent with the findings in [33], [53]. The improvements are significant for both i-vectors and x-vectors, especially at lower SNRs. As SNR increases, however, i-vectors based on GFCC performs worse than those based on MFCC in simulated RIRs after speech enhancement. But for x-vectors, GFCCs uniformly outperform MFCCs.

We also utilize single-channel WPE as a preprocessor to dereverberate the speech signal and report the results for real RIRs in Table III. Results show that further EER reduction is achieved when WPE is combined with speech enhancement.

B. Speaker Verification Based on Multi-Channel Speech Enhancement

Table IV presents EER results using the MVDR and rank-1 approximated MVDR beamformers for i-vectors and x-vectors. The results are for real RIRs, and both MFCC and GFCC are evaluated. Three different masks are derived from LSTM and GCRN: The estimated IRM obtained from LSTM (LSTM/IRM) and the IRM and PSM estimated from GCRN. The estimated masks computed from individual microphones are combined using median pooling. Following [38], we use

$$\mathbf{\Phi}_s(f) = \frac{\sum_t m(t,f)\boldsymbol{y}(t,f)\boldsymbol{y}(t,f)^H}{\sum_t m(t,f)} \tag{20}$$

to calculate the speech covariance matrix for rank-1 approximated MVDR. We find in our initial experiments that this derivation of the speech covariance matrix improves the results slightly compared to the derivation by Eq. (10).

From the table, we observe that the rank-1 approximation of the speech covariance matrix leads to a consistent improvement. Comparing the average improvements of the rank-1 approximated MVDR based on LSTM over the MVDR beamformer for the i-vector, we observe an absolute EER reduction of 5.23% with MFCC and 4.22% with GFCC. For x-vectors, rank-1 approximation leads to 4.72% and 2.88% absolute EER reduction with MFCC and GFCC features, respectively.

With the MVDR beamformer, our proposed approach for calculating covariance matrices based on GCRN outperforms that based on LSTM in all conditions for both i-vectors and x-vectors. Using GCRN/IRM with i-vectors, for example, the average improvements over LSTM/IRM are 3.11% and 2.39% EER reduction with MFCC and GFCC features, respectively.

For the rank-1 approximated MVDR beamformer, the use of GCRN brings an improvement over LSTM. GCRN/IRM and GCRN/PSM achieve lower EER scores compared to LSTM/IRM with both MFCC and GFCC features. Moreover, the PSM appears to be a more effective training target than the IRM for estimating covariance matrices as GCRN/PSM outperforms GCRN/IRM in all conditions. This is, however, reversed for the MVDR beamformer without rank-1 approximation, i.e. the IRM is a more effective training target in this case.

Table IV shows that, like the single-channel case, the use of GFCC improves speaker verification results uniformly over the use of MFCC. This is true regardless of DNN model, T-F mask, beamforming technique, SNR level, and speaker recognizer.

The performance of multi-channel speaker verification for simulated RIRs is presented in Table V. Similar trends to real RIRs are observed for both MFCC and GFCC features.

TABLE IV
MULTI-CHANNEL SPEAKER VERIFICATION RESULTS (%EER) BASED ON MFCC AND GFCC WITH REAL RIRS

		MFCC						GFCC					
		MVDR			MVDR Rank 1			MVDR			MVDR Rank 1		
		LSTM /IRM	GCRN /IRM	GCRN /PSM	LSTM /IRM	GCRN /IRM	GCRN /PSM	LSTM /IRM	GCRN /IRM	GCRN /PSM	LSTM /IRM	GCRN /IRM	GCRN /PSM
i-vector	0 dB	32.91	28.51	30.82	26.42	25.37	25.79	22.96	19.18	20.13	16.77	16.77	16.98
	5 dB	22.85	18.24	20.65	16.25	16.77	16.56	16.25	13.52	14.26	11.43	11.53	11.22
	10 dB	14.99	12.58	13.52	9.85	10.38	10.38	11.95	10.27	10.90	8.81	8.60	8.28
	15 dB	10.06	9.02	9.64	7.34	7.86	7.44	9.96	8.60	8.91	7.23	7.55	7.13
	Avg	20.20	17.09	18.66	14.97	15.10	15.04	15.28	12.89	13.55	11.06	11.11	10.90
x-vector	0 dB	24.21	21.17	22.75	17.40	17.40	16.56	20.75	16.98	18.45	14.68	14.36	14.05
	5 dB	16.67	13.84	15.20	10.69	11.11	10.27	12.58	10.48	11.53	9.43	8.70	8.49
	10 dB	11.74	9.54	10.59	7.86	7.55	7.34	8.39	7.23	7.86	6.81	5.87	5.77
	15 dB	8.49	7.13	7.44	6.29	6.18	5.98	6.39	6.18	6.29	5.66	5.14	5.03
	Avg	15.28	12.92	14.00	10.56	10.56	10.04	12.03	10.22	11.03	9.15	8.52	8.34

TABLE V
MULTI-CHANNEL SPEAKER VERIFICATION RESULTS (%EER) BASED ON MFCC AND GFCC WITH SIMULATED RIRS

		MFCC						GFCC					
		MVDR			MVDR Rank 1			MVDR			MVDR Rank 1		
		LSTM /IRM	GCRN /IRM	GCRN /PSM	LSTM /IRM	GCRN /IRM	GCRN /PSM	LSTM /IRM	GCRN /IRM	GCRN /PSM	LSTM /IRM	GCRN /IRM	GCRN /PSM
i-vector	0 dB	9.33	7.44	7.97	6.60	6.39	6.50	7.55	6.81	6.71	6.60	6.60	6.60
	5 dB	5.24	4.40	4.51	4.19	4.09	4.09	5.24	4.82	4.82	4.82	4.82	5.03
	10 dB	3.98	3.67	3.67	3.77	3.35	3.67	4.30	4.19	4.19	4.40	4.19	4.30
	15 dB	3.67	3.56	3.46	3.35	3.14	3.56	3.77	4.09	4.09	4.19	4.09	4.30
	Avg	5.55	4.77	4.90	4.48	4.24	4.45	5.22	4.98	4.95	5.00	4.93	5.06
x-vector	0 dB	5.03	4.19	4.19	4.40	3.88	3.77	4.51	3.98	3.98	4.09	3.88	3.98
	5 dB	3.67	3.35	3.35	3.14	3.14	3.14	2.94	2.94	2.94	2.83	3.04	3.25
	10 dB	2.72	2.72	2.72	2.83	2.52	2.72	2.83	2.73	2.62	2.62	2.62	2.83
	15 dB	2.72	2.52	2.52	2.52	2.72	2.83	2.73	2.52	2.41	2.41	2.52	2.94
	Avg	3.53	3.19	3.19	3.22	3.06	3.11	3.25	3.04	2.99	2.99	3.02	3.25

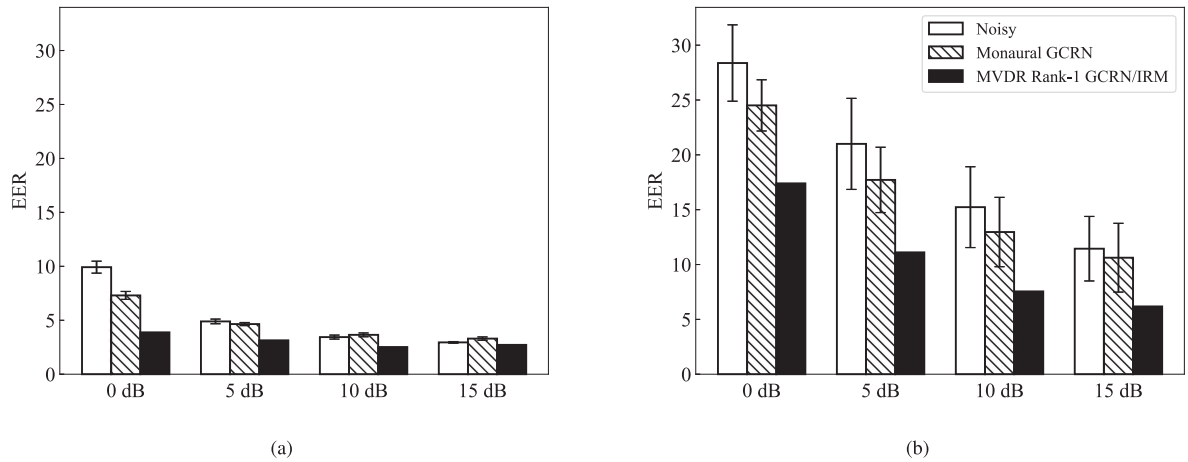

Fig. 5. EER results for single- and multi-channel speech enhancement for x-vectors based on MFCC features with (a) simulated RIRs and (b) real RIRs. For noisy and monaural GCRN, the height of a bar and the whisker denote the average EER and the standard deviation of all microphones, respectively.

Analogous to single-channel experiments, GFCC performance is worse than MFCC for i-vectors as SNR increases.

To gain an insight into the effect of multi-channel speech enhancement, Fig. 5 contrasts the EER score of the rank-1 approximated MVDR beamformer based on GCRN/IRM with the unprocessed audio of each microphone and monaurally enhanced speech signal obtained by GCRN. One can clearly observe that greater error reduction is achieved by using the rank-1 approximated MVDR beamformer compared to monaural speech enhancement, especially at lower SNRs. Unlike

TABLE VI
MULTI-CHANNEL SPEAKER VERIFICATION RESULTS (%EER) USING WPE PREPROCESSING WITH REAL RIRS. 'BF' REFERS TO RANK-1 APPROXIMATED MVDR BEAMFORMING WITH GCRN BASED MASKING

		MFCC			GFCC		
	SNR	WPE	WPE +BF/IRM	WPE +BF/PSM	WPE	WPE +BF/IRM	WPE +BF/PSM
i-vector	0 dB	42.14	20.96	20.96	32.60	13.63	13.42
	5 dB	37.63	12.89	13.00	26.00	8.49	8.39
	10 dB	29.56	7.86	7.55	20.13	6.60	6.50
	15 dB	20.44	5.24	5.14	15.41	5.77	5.45
	Avg	32.44	11.74	11.66	23.54	8.62	8.44
x-vector	0 dB	35.01	13.21	12.89	29.14	10.90	11.01
	5 dB	26.83	8.29	7.76	22.22	7.02	7.02
	10 dB	20.96	6.08	5.77	16.14	4.40	4.61
	15 dB	16.35	4.72	4.51	12.16	3.77	3.88
	Avg	24.79	8.07	7.73	19.92	6.53	6.63

TABLE VII
MULTI-CHANNEL EER COMPARISON EVALUATED ON NIST RETRANSMITTED DATASET WITH REAL RIRS

	i-vector		x-vector	
	MFCC	GFCC	MFCC	GFCC
FW_GEV [23]	7.54	–	–	–
MVDR Rank 1 GCRN/IRM	6.08	7.23	5.24	4.51
MVDR Rank 1 GCRN/PSM	5.66	7.12	4.82	**4.19**
WPE	5.76	7.65	4.82	5.03
WPE + FW_GEV [23]	2.73	–	–	–
WPE + MVDR Rank 1 GCRN/IRM	2.93	4.82	3.24	**2.62**
WPE + MVDR Rank 1 GCRN/PSM	3.24	4.40	3.14	2.93

monaural speech enhancement which causes some degradation on x-vector performance in the simulated RIRs at higher SNRs, multi-channel speech enhancement improves speaker verification performance of the x-vector systems uniformly.

Table VI shows the application of multi-channel WPE for preprocessing with and without rank-1 approximated MVDR beamforming for real RIRs. Comparing this table and Table IV, it can be seen that the combination of WPE and beamforming improves the performance significantly, especially at lower SNRs.

Finally, we compare our multi-channel SR system to the system introduced in [23]. To match the setup in [23], this comparison uses the real RIR corpus without adding babble noise. We report the results in Table VII. FW_GEV refers to the masking-based generalized eigenvalue beamformer [14] that is trained on reverberant data. The training procedure for i-vectors and PLDA used in [23] is similar to our procedure. For i-vectors, our system based on MVDR rank-1 GCRN/PSM and MFCC features yields 1.88% lower EER than [23]. We improve EER scores further by using x-vectors, which is not used in [23]. It is also worth noting that rank-1 approximated MVDR with the PSM has better EER scores compared to WPE, indicating the beamformer's ability to effectively dereverberate. Even better EER scores are achieved when WPE is combined with masking-based beamformers. The best EER is achieved by MVDR rank-1 GCRN/IRM using x-vectors and GFCC features.

VII. CONCLUSIONS

In this paper, we have investigated the effectiveness of speech enhancement methods for speaker recognition in adverse acoustic conditions where the target speech is corrupted by strong diffuse noise and room reverberation. Convolutional recurrent networks are used to perform monaural speech enhancement, and they learn to map from the real and imaginary spectrograms of noisy speech to the corresponding spectrograms of clean speech. Our experimental results demonstrate that complex spectral mapping with GCRN combined with GFCC features significantly reduces speaker verification errors of both i-vector and x-vector systems. Our experiments show that multi-channel speech enhancement consistently improves the x-vector performance. Two variants of the masking-based MVDR beamformer have been investigated for multi-channel speech enhancement. The results demonstrate that rank-1 approximation leads a greater error reduction, likely due to more accurate steering vector estimation.

ACKNOWLEDGMENT

The authors would like to thank Ke Tan for helpful discussions, and L. Mošner and J. Černockỳ for providing data and discussions regarding [23].

REFERENCES

[1] "NIST, 2008 speaker recognition evaluation plan," 2008. [Online].Available: https://www.nist.gov/sites/default/files/documents/2017/09/26/sre08_ev alplan_release4.pdf
[2] J. Chang and D. L. Wang, "Robust speaker recognition based on DNN/i-vectors and speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5415–5419.
[3] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, pp. 2604–2612, 2016.
[4] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 933–941.
[5] N. Dehak, P. J. 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] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 708–712.
[7] L. Ferrer *et al.*, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proc. NIST Workshop*, 2011, pp. 1–7.
[8] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
[9] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.
[10] S. Gannot, E. Vincent, S. M.-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
[11] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, pp. 2911–2917, 2008.
[12] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015.

[13] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5115–5119.

[14] J. Heymann, L. Drude, and R. H.-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 196–200.

[15] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5210–5214.

[16] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. Eur. Conf. Comput. Vision*, 2006, pp. 531–542.

[17] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4253–4256.

[18] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1695–1699.

[19] N. Li and M.-W. Mak, "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1648–1659, Oct. 2015.

[20] A. Lozano-Diez *et al.*, "Analysis and optimization of bottleneck features for speaker recognition." in *Proc. Odyssey Speaker, Lang. Recognit. Workshop*, 2016, pp. 21–24.

[21] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4814–4818.

[22] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. Interspeech*, 2017, pp. 2008–2012.

[23] L. Mošner, P. Matějka, O. Novotnỳ, and J. H. Černockỳ, "Dereverberation and beamforming in far-field speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5254–5258.

[24] O. Novotnỳ *et al.*, "Analysis of DNN speech signal enhancement for robust speaker recognition," *Comput. Speech Lang.*, vol. 58, pp. 403–421, 2019.

[25] O. Novotnỳ, O. Plchot, P. Matejka, L. Mosner, and O. Glembek, "On the use of x-vectors for robust speaker recognition." in *Proc. Odyssey Speaker, Lang. Recognit. Workshop*, 2018, pp. 168–175.

[26] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, pp. 465–494, 2011.

[27] O. Plchot, L. Burget, H. Aronowitz, and P. Matejka, "Audio enhancing with DNN autoencoder for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5090–5094.

[28] D. Povey *et al.*, "The Kaldi Speech Recognition Toolkit," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011, pp. 1–4.

[29] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *Int. Conf. Learn. Representations*, 2018.

[30] J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matějka, and L. Burget, "End-to-end DNN based speaker recognition inspired by i-vector and PLDA," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4874–4878.

[31] S. O. Sadjadi and J. H. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. Interspeech*, 2010, pp. 2138–2142.

[32] R. Serizel, M. Moonen, B. V. Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in Cochlear implants," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[33] Y. Shao and D. L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 1589–1592.

[34] S. Shon, H. Tang, and J. Glass, "VoiceID loss: Speech enhancement for speaker verification," in *Proc. Interspeech*, 2019, pp. 2888–2892.

[35] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484v1*.

[36] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 92–97.

[37] D. Snyder, D. G.-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.

[38] X. Sun, Z. Wang, R. Xia, J. Li, and Y. Yan, "Effect of steering vector estimation on MVDR beamformer for noisy speech recognition," in *Proc. IEEE 23rd Int. Conf. Digital Signal Process.*, 2018, pp. 1–5.

[39] H. Taherian, Z.-Q. Wang, and D. L. Wang, "Deep learning based multi-channel speaker recognition in noisy and reverberant environments," in *Proc. Interspeech*, 2019, pp. 4070–4074.

[40] K. Tan and D. L. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6865–6869.

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

[42] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[43] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[44] Z.-Q. Wang and D. L. Wang, "Integrating spectral and spatial features for multi-channel speaker separation," in *Proc. Interspeech*, 2018, pp. 2718–2722.

[45] Z.-Q. Wang and D. L. Wang, "Mask weighted STFT ratios for relative transfer function estimation and its application to robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5619–5623.

[46] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Comput. Speech Lang.*, vol. 49, pp. 37–51, 2018.

[47] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[48] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6745–6749.

[49] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 436–443.

[50] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[51] X. Zhang, Z.-Q. Wang, and D. L. Wang, "A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 276–280.

[52] F. Zhao, H. Li, and X. Zhang, "A robust text-independent speaker verification method based on speech separation and deep speaker," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6101–6105.

[53] X. Zhao, Y. Shao, and D. L. Wang, "CASA-based robust speaker identification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1608–1616, Jul. 2012.

[54] X. Zhao and D. L. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7204–7208.