Compensating noise and reverberation in far-field Multichannel Speaker Verification

Sandipana Dowerah, Romain Serizel, Denis Jouvet, Mohammad Mohammadamini, Driss Matrouf

Compensating noise and reverberation in far-field Multichannel Speaker Verification

1st Sandipana Dowerah
Université de Lorraine, CNRS, Inria, Loria
Nancy, France
sandipana.dowerah@loria.fr

2nd Romain Serizel
Université de Lorraine, CNRS, Inria, Loria
Nancy, France
romain.serizel@loria.fr

3rd Denis Jouvet
Université de Lorraine, CNRS, Inria, Loria
Nancy, France
denis.jouvet@loria.fr

4th Mohammad Mohammadamini
Avignon Université, Laboratoire Informatique d'Avignon,
Avignon, France
mohammad.mohammadamini@univ-avignon.fr

5th Driss Matrouf
Avignon Université, Laboratoire Informatique d'Avignon,
Avignon, France
driss.matrouf@univ-avignon.fr

***Abstract*—Speaker verification (SV) suffers from unsatisfactory performance in far-field scenarios due to environmental noise and the adverse impact of room reverberation. This paper investigates utilizing a multichannel pre-processing pipeline including a time-domain neural beamformer (FaSNet), multichannel Wiener filter (MWF), and weighted prediction error (WPE). This approach is compared to the existing state-of-the-art approaches. We examine the importance of enrollment in pre-processing which has been largely overlooked in previous studies. Experimental evaluation shows that pre-processing can improve the SV performance as long as the enrollment files are processed similarly to the test data and that test and enrollment occur within similar SNR ranges. The integration of FaSNet, MWF, and WPE achieved improved performance compared to the existing state-of-the-art pre-processing approaches. We also show that our approach generalizes to unseen real recorded data while being trained on simulated data.**

***Index Terms*—multichannel speech enhancement, far-field speaker verification**

I. Introduction

Speaker verification (SV) is the process of analysing the authenticity of a speaker on the basis of his/her voice characteristics. SV is becoming an integral part to avail services in many sectors like banking, online payment systems, etc. However, the real-world is a noisy one and the efficacy of SV system under far-field setting is still a challenging task. This is mainly due to distortion of the original speech signal as effects of the long range fading, room reverberation and complex environmental noises. To address this problem, several challenges has been organised over the past few years such as, VOiCES from a distance challenge [1], Interspeech far-field speaker verification challenge [2], etc. The current state-of-the-art x-vector [3] based approaches improved the SV performance significantly. But, these SV systems still suffer from severe performance degradation in noisy-reverberant scenarios that are typical of hands-free applications.

Speech enhancement can improve the overall quality of a degraded speech signal. Besides denoising autoencoder [4], [5], neural beamforming [6] and dereverberation [7] has been extensively used as front-end processing of speech recognition [6], [12], [28]. But, only few studies have examined the effectiveness of integrating beamforming and dereverberation with multichannel signal for SV in noisy-reverberant environment [8], [9]. Mosner et al. employed mask-based beamforming combined with WPE to minimize the reverberation effect but they studied only reverberation effect whereas reverberation and noise occurs simultaneously in real scenarios [8]. Yang et al. jointly optimized neural network supported minimum variance distortionless response (MVDR) beamforming with WPE using deep speaker embedding model [10]. Taherain et al. used MVDR beamformer with Rank-1 approximation to search for the optimal beamformer from the variants of ideal ratio mask based MVDR and generalized eigenvalue (GEV) beamformers [9]. Although often used in a multichannel context, most of these studies use single-channel data as an input to DNN, use matched train/test data, and reported poor performance on real data. Moreover, prior works mainly used mask-based beamformers (MVDR or GEV) in frequency domain which typically degrades in causal and online scenarios [13] as frequency domain methods lacks the reasonable size of frequency resolution and input signal length required for perceivable system latency.

This paper investigates multichannel pre-processing for SV in adverse acoustic condition where noise and room reverberation distorts the target speech signal. We integrate a time-domain neural beamformer, Rank-1 MWF and WPE as a multichannel pre-processing to SV. We employ FaSNet to compute the time-frequency (T-F) masks which inherently considers phase information as well. Further, we examine the importance of enrollment in pre-processing as different enrollment/test mismatches can have different impact on SV performance. Our

French National Research Agency supports this work for ROBOVOX project (ANR-18-CE33-0014)). Experiments were partially carried out using the Grid5000 testbed supported by a scientific group of Inria including CNRS, RENATER and other Universities and organizations (see https://www.grid5000) hosted by the University of Lorraine.

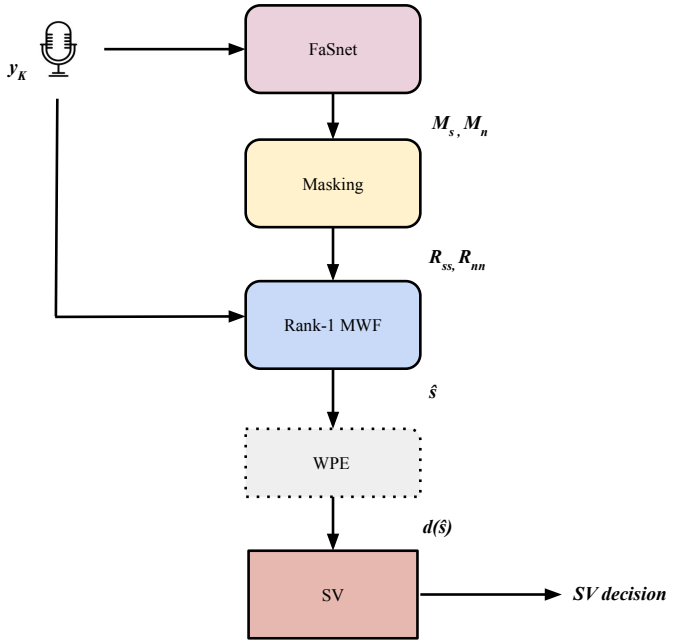

Fig. 1: *Proposed multichannel speech enhancement pipeline. y is the noisy input from the K channels. M_S, M_n are the computed masks for speech and noise. R_{ss}, R_{nn} are speech and noise covariance matrices, $\hat{s}$ is the enhanced speech and $d(\hat{s})$ is the dereverb enhanced speech .*

proposed system employs Rank-1 MWF which is robust to low SNR scenarios and provides noise reduction extensively. We also investigate the influence of quality (in terms of source to distortion and source to interference ratio) of the enhanced signals which could be helpful in fine-tuning the front-end of a SV system.

II. Problem Formulation

A. Signal Model

Considering the mixture of dry speech and noise as recorded by K microphones can be formulated with the short-time Fourier transform (STFT) as $y(T-F) = s(T-F) + h(T-F) + n(T-F)$, where $y(T-F)$, $s(T-F)$ $h(T-F)$ and $n(T-F)$ represent the STFT vectors of the noisy speech, dry speech, reverberated speech and noise.

III. Multichannel Speech Enhancement

A. FaSNet

FaSNet is a time-domain neural beamforming algorithm to separate noise and speech [13]. We use FaSNet to compute the T-F masks:

$$M_s(T-F) = \frac{|s(T-F)|}{|s(T-F)| + max(|n(T-F), \varepsilon)} \quad (1)$$

$$M_n(T-F) = \frac{|n(T-F)|}{|s(T-F)| + max(|n(T-F)|, \varepsilon)} \quad (2)$$

Where, ε is 1×10^{-16}.

B. Rank-1 MWF

MWF is designed to minimize mean squared error (MSE) criterion between the record mixture and the target speech.

$$J(\boldsymbol{w}) = \mathbb{E}\{|s_1 - \mathbf{w}^H \mathbf{y}|^2\} \quad (3)$$

where s_1 is the clean speech signal from the first channel, $\mathbb{E}\{.\}$ is the expectation operator and $\cdot^H$ denotes the Hermitian transpose. The filter $\mathbf{w}$ that minimizes the MSE criterion (3) is the MWF that can be expressed as below:

$$\hat{\mathbf{w}}_{\mathrm{MWF}}(f) = (\mathbf{R}_{ss}(f) + \mathbf{R}_{nn}(f))^{-1} \mathbf{R}_{ss}(f) \mathbf{u}_1 \quad (4)$$

Where, $R_{ss}(f)$, $R_{nn}(f)$ are spatial correlation matrix for the speech and noise, respectively and $\mathbf{u}_1 = [1, \ldots, 0]^T$.

It is possible to introduce a trade-off parameter μ which controls the tradeoff between the interference reduction and the desired signal distortion [14]. We then obtained the speech distortion weighted (SDW) MWF that can be expressed as:

$$\hat{\mathbf{w}}_{\mathrm{SDW-MWF}}(f) = (\mathbf{R}_{ss}(f) + \mu \mathbf{R}_{nn}(f))^{-1} \mathbf{R}_{ss}(f) \mathbf{u}_1 \quad (5)$$

If the desired signal comes from a single source, the speech correlation matrix $\mathbf{R}_{ss}$ is theoretically of Rank-1. Forcing this matrix to its Rank-1 approximation leads to the so-called Rank-1 version of the filters described above. In the remainder of the paper, we use Rank-1 approximation of the SDW-MWF.

The computation of MWF requires estimation of the speech and noise correlation matrices. The estimated T-F masks of speech and noise are used to compute the spatial correlation matrices $\mathbf{R}_{ss}(f)$ and $\mathbf{R}_{nn}(f)$ that are needed to derive the MWF. The correlation matrices are obtained as:

$$\mathbf{R}_{ss}(f) = \frac{1}{T} \sum_{t=0}^{T-1} \check{\mathbf{s}}(T-F) \check{\mathbf{s}}(T-F)^H \quad (6)$$

Note that the noise correlation matrix can be obtained similarly as in Eq. (6).

C. WPE

WPE is used for alleviating degradation performance in speech recognition mostly in the case of a far-field scenario. The de-reverberated signal is obtained by subtracting the filtered signal from the observed signal denoted by;

$$d(\hat{s}) = \hat{s}(t) - \Sigma_{k=1}^{N} \hat{w}(k) h(t-k) \quad (7)$$

Where, $\hat{s}(t)$ is reverberated signal at time t and $d(\hat{s})$ is de-reverberated signal using WPE algorithm. $\hat{w}$ denotes the k^{th} tap of the N-taps. WPE filter is $W = [W_1, \ldots, W_N]^T$.

IV. Dataset

A. Synthetic dataset

We generated a synthetic dataset namely, RoboVoices simulating real room environments with additive noise and reverberation from dry speech segments. Designing such a dataset is necessary as the training of speech enhancement approaches requires ground-truth knowledge about the target speech and to some extent about the degradation. This kind of information is not available in the available corpora for far-field SV.

1) Speech data: We use the dry speech data from the clean subset of Librispeech [15] corpus which is approximately 1000 hours of English speech data collected as part of the Librivox project. We have selected around 10000 files randomly from the dry training subset of Librispeech and truncated them to 10 seconds duration for the training set, contributing to 25 hours of speech data.

For evaluation of the SV system, we use the Fabiole speech corpus [18]. Fabiole is a French speech corpus consisting of around 6882 audio files from 130 native French speakers. The minimum duration of speech file is 1 seconds and the maximum is 46 seconds. The speech data of the corpus is collected from different French radio and television shows. For creating each evaluation set, we have used 1200 speech files from Fabiole representing 2 hrs of evaluation material.

2) Noise data: We have collected realistic office noise from the Freesound platform[1] [16]. The selected noise categories include door, keyboard, office, phone, background noise in the room, printer, fan, door knock, babble, and environmental noise, etc. We split the dataset into a training set comprised of 3725 clips and an evaluation set comprised of 1000 clips.

We also evaluate our system's performance using MUSAN noise from the OpenSRL dataset [2] [21]. We convolved the dry speech from Librispeech and noise from Musan with simulated RIR for training. The evaluation protocol is same as RoboVoices except the noise samples. The noise categories include dial tones, raindrops, etc.

3) Room Impulse Response: To simulate room effects, we have generated a RIR corpus of 10000 rooms for training and 3600 for evaluation with pyroomacoustics toolbox [17]. For training, the room length was chosen between $[3-8]$ m, width was chosen between $[3-5]$ m, and the height was chosen between $[2-3]$ m. The absorption coefficient was drawn randomly such that the room's RT60 was between $[200-600]$ ms. The minimum distance of a source and the wall is 1.5 m and 1 m between the wall and the microphones. The RIR for the evaluation set were generated with the same room dimension as in the training set but the absorption coefficient was selected to obtain an RT60 of 400 ms.

The final RoboVoices corpus for training and evaluation is created by first convolving the dry speech and noise with the simulated RIRs. We then added the convolved dry speech and convolved noise to obtain the noisy signal. We randomly select the noise samples from Freesound and the dry speech from Librispeech for the training set. The SNR is drawn randomly with a uniform distribution between $[0-10]$ dB. For the evaluation set, the generation process is similar except that we draw the SNR values in $[5, 10, 20]$dB and the process is applied to each speech segment from the Fabiole dataset. In total, we have generated 10000 mixture for training and 3600 mixture for evaluation.

[1]https://freesound.org/

[2]https://www.openslr.org/index.html

TABLE I: *Results on Freesound noise and Musan noise using different pre-processing methods. FaS is FaSNet in the table.*

Noise type	Freesound			Musan		
Pre-processing/SNR	**SDR**	**SIR**	**EER**	**SDR**	**SIR**	**EER**
dry speech	—	—	14.9	—	—	14.9
Reverb-speech	—	—	20.6	—	—	20.6
Noisy	2.6	15.1	28.2	2.4	14.8	25.7
BLSTM GEV-BAN	5.4	20.6	27.1	4.3	20.2	23.8
BLSTM Rank-1	5.4	20.7	26.8	5.1	20.7	23.1
BLSTM MVDR Rank-1	5.8	20.1	27.0	4.9	20.5	23.7
FaS	5.3	20.5	38.7	4.7	18.3	32.6
FaS GEV-BAN	5.8	21,9	26.8	5.6	21.2	22.2
FaS Rank-1 MWF	6.1	21.0	24.9	5.9	21.5	21.9
FaS Rank-1 MWF WPE	**7.0**	**21.0**	**23.3**	**6.1**	**21.5**	**20.5**

B. VOiCES

Additionally, we evaluate our approach on the VOiCES challenge 2019 dataset [1]. Among 11 microphone positions in the Eval set, we select 3 representative positions: 2, 4, and 9. We select the signal from these three microphones confirming all three are in mid-distance from the speaker and are close to build a "virtual" microphone antenna.

V. Experimentation

A. Experimental Set-up

The speech and noise signals are sampled at 16 kHz. We provide multichannel speech signal as input to FaSNet with 4 ms window size and context size of 16 ms. We trained the FaSNet model with SDR loss and SI-SNR (scale-invariant source-to-noise ratio) loss [19]. We employed the dual-path RNN (DPRNN) with an encoder dimension of 50, a chunk size of 50, and a hopping window of 35 dimension. To compute the target masks, we use the source-separated outputs from the FaSNet model. The FaSNet implementation is used from Asteroid toolbox [20] and replaced the TCN blocks with DPRNN in contrast to the original FaSNet architecture, where TCN is used to predict the beamformed filters.

The SDW-MWF operate on T-F representation of the signal. STFT is computed with a window length of 512 samples, a hop size of 256 samples and a Hann window. A single SDW-MWF is estimated for each speech clips. According to previous experiments we set μ parameter of the SDW-MWF to 0.1 to limit the amount of distortion introduced by the filter. We use WPE with the following parameters: 10 filter taps, a delay of 3 frames, 5 iterations of WPE and alpha 0.9999.

B. Speaker Verification

Our SV is an x-vector [3] based system. The network is trained with data augmentation using different portions of Musan corpus (music, babble, noise, reverberation) [21] with 1 million augmented files from Voxceleb [22] and all the original files from Voxceleb 1 and 2 [23]. We use Fabiole corpus for tests and enrollment. For enrollment, 3441 files are used and the remaining files are used for the test. As input to the x-vector network, we extract Mel-frequency cepstral coefficients normalized by Cepstral Mean-Variance Normalization. The non-speech frames are removed with a

TABLE II: *% EER on matched pre-processing conditions on the RoboVoices dataset. We processed both enrollment and test data using the same range of SNR. The average confidence interval is 0.1.*

	Enrollment conditions				
Test data	Dry speech	Reverb. speech	Noisy	BLSTM MVDR Rank-1	FaS Rank-1 MWF WPE
Dry speech	**14.9**	15.4	16.7	16.1	15.7
Reverb-speech	20.6	**19.8**	20.5	20.4	20.1
Noisy	28.2	24.9	**23.8**	24,9	24,3
BLSTM MVDR Rank-1	27.0	24.2	23.4	**21.3**	22.5
FaS Rank-1 MWF WPE	23.3	22.8	21.5	22.4	**19.2**

TABLE III: *%EER on RoboVoices using different pre-processing methods.*

Noise type	**RoboVoices**		
Pre-proces./SNR	5	10	20
Noisy	34.4	28.0	22.2
BLSTM GEV-BAN	32.5	26.8	21.9
BLSTM MVDR Rank-1	32.3	26.6	22.1
FaS Rank-1 MWF WPE	**27.1**	**23.2**	**19.7**

TABLE IV: *%EER on different noise conditions of the VOiCES Eval dataset. Confidence interval is 0.2. Reverb is reverberated.*

	Noise conditions			
SV pre-processing	Clean	Babble	TV	Music
None	4.4	9.2	7.9	8.4
BLSTM Rank-1	4.4	8.1	7.1	7.3
BLSTM MVDR Rank-1	4.3	7.3	6.5	6.9
FaS	4.4	7.8	7.4	7.9
FaS Rank-1 MWF	4.5	7.1	6.8	7.1
FaS Rank-1 MWF WPE	**4.0**	**6.3**	**6.0**	**6.4**

voice activity detector. The Probabilistic Linear Discriminant Analysis (PLDA) classifier used for scoring is trained on 200k x-vectors extracted from Voxceleb. Before training the PLDA, x-vectors are centered and their dimensionality reduced to 128 with linear discriminant analysis. The PLDA scoring system is retrained on the enrollment set. Kaldi [27] [3] is used to process all the steps of SV.

C. Evaluation

Speech enhancement results were evaluated in terms of source-to-distortion (SDR) ratio for estimating distortion on the target signal and the source to interference (SIR) ratio for estimating the relative importance of the estimated target speech compared to uncorrelated interference [24] [4]. The SV system is evaluated using an equal error rate (EER). All metrics are presented with a 95 % confidence interval using the bootstrap algorithm [25]. We compute EER on dry speech and reverberated speech (as a reference point), on the input mixture, and on the signals estimated with different speech enhancement algorithms.

VI. Results and Analysis

Table I presents the results of different state-of-the-art pre-processing techniques on Freesound and Musan noise datasets. We implement the BLSTM-based approaches from [9] and consider them as the baseline. The performance is averaged over SNR conditions. The enrollment is always done using dry speech here. We can see from the table that both reverberation and noise degrades the SV performance. SDR seems to be closely co-related with EER but not in the case of SIR. Comparing the average improvements of FaS Rank-1 MWF WPE to Rank-1 approximated MVDR, we observe an absolute EER reduction of 5% on both Fressound and Musan noise datasets. In terms of EER, FaSNet is outperformed significantly by Rank-1-based beamforming approaches. This could be due to artifacts introduced by FaSNet as indicated by lower SDR values. The integration of FaSNet, Rank-1 MWF, and WPE bring substantial improvement over all other techniques irrespective of the noise types present in both datasets. RoboVoices has non-stationary noises which explain the high EER compared to Musan. FaS Rank-1 MWF WPE surpasses all the BLSTM-based baseline approaches on both datasets. We will be using only the best baseline approach i.e BLSTM MVDR Rank-1 and our proposed FaS Rank-1 WPE in the forthcoming experiments.

Table II reports the performance on the RoboVoices dataset for different pre-processing conditions and depending on the enrollment condition. Performing the enrollment and test with matched acoustic conditions alleviates the effect of reverberation but this is hardly the case for additive noise. Pre-processing consistently improves the SV performance but the effectiveness is more evident when the enrollment is done in matched pre-processing conditions (diagonal). FaS Rank-1 MWF WPE obtained the best EER performance for a noisy and reverberated input over the baseline approach.

Table III shows the performance of different pre-processing approaches depending on SNR conditions on the RoboVoices dataset. Comparing the consistent improvement of FaS Rank-1 MWF WPE with BLSTM GEV-BAN and Rank-1 approximated BLSTM MVDR, we observe an absolute reduction of EER across the SNR conditions for a noisy input signal. With 7% EER reduction at 5 dB, FaS Rank-1 MWF WPE shows robustness to low SNR conditions. Thus, supporting the argument that Rank-1 MWF is robust to low SNR scenarios.

Table IV presents the results obtained for various distractor noise conditions on the VOiCES Eval dataset. We select the microphone which was closest to the speaker as a reference microphone. As expected, the condition with no noise distractor (Clean in Table IV) resulted in the best performance across

[3] https://github.com/kaldi-asr/kaldi

[4] SDR and SIR are computed with the mir_eval toolbox https://github.com/craffel/mir_eval

all the approaches. The baseline BLSTM-based approaches performs poorly compared to FaSNet-based approaches in all the noise conditions. With an EER of 9.2% without any pre-processing, Babble seems to be the most challenging condition due to overlapping speech interference as well as its similarity to the desired clean speech. However using FaS Rank-1 MWF WPE, EER of Babble reduces to 6.3%. Furthermore, FaS Rank-1 MWF WPE achieves the best performance across the noise conditions demonstrating the efficacy of our approach even though the model was trained on synthetic data generated for generic, possibly mismatched, and spatial scenarios. Notably, it shows that our approach generalizes to unseen noise such as Babble for which the performance was improved using FaS Rank-1 MWF WPE as a pre-processing pipeline.

VII. Conclusion

This paper demonstrates the efficacy of integrating a neural beamformer, Rank-1 MWF and WPE as a pre-processing for speaker verification in multichannel distant/far-field audio under noisy-reverberant conditions. The proposed approach outperformed the existing state-of-the-art approaches in terms of EER. Experimentation with enrollment shows that performing the test and enrollment with matched acoustic conditions alleviates the effect of reverberation. Additionally, our approach shows more robustness to low SNR conditions. The integration of FaS Rank-1 MWF WPE as a pre-processing demonstrated the best performance across the noise conditions on the VOiCES challenge dataset even though the model was trained on synthetic data. This shows that our approach generalizes to unseen real recorded data.

References

[1] Mahesh Kumar Nandwana and Julien van Hout and Colleen Richey and Mitchell McLaren and M. Barrios and A. Lawson, "The VOiCES from a Distance Challenge 2019, INTERSPEECH", 2019. .

[2] Xiaoyi Qin, Ming Li, Hui Bu, Wei Rao, Rohan Kumar Das, Shrikanth S. Narayanan, and Haizhou Li, "The interspeech 2020 far-field speaker verification challenge," ArXiv, vol. abs/2005.08046, 2020.

[3] David Snyder, D. Garcia-Romero, Gregory Sell, Daniel Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[4] Oldrich Plchot, L. Burget, H. Aronowitz, and P. Matejka, "Audio enhancing with dnn autoencoder for speaker recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2016.

[5] Cheng Yu, Ryandhimas E. Zezario, Syu-Siang Wang, Jonathan Sherman, Yi-Yen Hsieh, Xugang Lu, Hsin-Min Wang, and Yu Tsao, "Speech enhancement based on denoising autoencoder with multi-branched encoders," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 2020.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in Proc. Int. Conf. Acoust., Speech, Signal Process., 2016.

[7] Y. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 20, 2012.

[8] L. Mosner, P. Matejka, O. Novotny, and J. H. Cernocky, "Dereverberation and beamforming in far-field speaker recognition," in Proc. Int. Conf. Acoust., Speech, Signal Process., 2018.

[9] Hassan Taherian, Zhong-Qiu Wang, DeLiang Wang, "Deep Learning Based Multi-channel Speaker Recognition in Noisy and Reverberant Environments", INTERSPEECH, 2019.

[10] Joon-Young Yang, Joon-Hyuk Chang, "Joint optimization of neural acoustic beamforming and dereverberation with x-vectors for robust speaker verification", INTERSPEECH 2019.

[11] Hassan Taherian, Zhong-Qiu Wang, DeLiang Wang, "Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement", IEEE Trans. on Audio, Speech and Language Processing, vol-28, 2020.

[12] Christoph Boeddeker, Hakan Erdogan, Takuya Yoshioka, and Reinhold Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018

[13] Yi Luo, Enea Ceolini, Cong Han, Shih-Chii Liu, and Nima Mesgarani, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019.
to action segmentation," ArXiv, vol. abs/1608.08242, 2016.

[14] Shmulik Markovich Golan, Sharon Gannot, and Israel Cohen, "Performance of the sdw-mwf with randomly located microphones in a reverberant enclosure," IEEE Transactions on Audio, Speech, and Language Processing, vol.21, 2013.

[15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[16] Eduardo Fonseca, Jordi Pons, Xavier Favory, F. Font, D. Bogdanov, Andr es Ferraro, Sergio Oramas, Alastair Porter, and X. Serra, "Freesound datasets: A platform for the creation of open audio datasets," in ISMIR, 2017

[17] Robin Scheibler, Eric Bezzam, and Ivan Dokmani c, "Pyroomacoustics: A python package for audio room simula- tion and array processing algorithms," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[18] M. Ajili, J. Bonastre, Juliette Kahn, S. Rossato, and Guillaume Bernard, "Fabiole, a speech database for forensic speaker comparison," in LREC, 2016.

[19] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R.Hershey, "Sdr– half-baked or well done?," 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[20] Manuel Pariente, S. Cornell, Joris Cosentino, S. Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian Robert Stoter, Mathieu Hu, Juan M. Mart'in-Donas, David Ditter, Ariel Frank, Antoine Deleforge, and E. Vincent, "Asteroid: the pytorch-based audio source separation toolkit for researchers," ArXiv, vol. abs/2005.04132, 2020.

[21] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," ArXiv, vol. abs/1510.08484, 2015.

[22] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in INTERSPEECH, 2017.

[23] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in INTERSPEECH, 2018.

[24] E. Vincent, R. Gribonval, and C. F evotte, "Performance measurement in blind audio source separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, 2006.

[25] M. Bisani, H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 2004.

[26] Simon Doclo and Marc Moonen, "Gsvd-based optimal filtering for single and multi microphone speech enhancement," IEEE Trans. Signal Process., vol. 50, 2002.

[27] Povey, Daniel Ghoshal, Arnab Boulianne, Gilles Burget, Lukáš Glembek, Ondrej Goel, Nagendra Hannemann, Mirko Motlíček, Petr Qian, Yanmin Schwarz, Petr Silovský, Jan Stemmer, Georg Vesel, Karel, "The Kaldi speech recognition toolkit", IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2011.

[28] K. Kinoshita, M. Delcroix, H. Kwon, T. Hori, and T. Nakatani, "Neural network based spectrum estimation for online WPE dereverberation," in Proc. Interspeech, 2017.