DEREVERBERATION AND BEAMFORMING IN FAR-FIELD SPEAKER RECOGNITION

Ladislav Mošner, Pavel Matějka, Ondřej Novotný and Jan “Honza” Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia

ABSTRACT

This paper deals with far-field speaker recognition. On a corpus of NIST SRE 2010 data retransmitted in a real room with multiple microphones, we first demonstrate how room acoustics cause significant degradation of state-of-the-art i-vector based speaker recognition system. We then investigate several techniques to improve the performances ranging from probabilistic linear discriminant analysis (PLDA) re-training, through dereverberation, to beamforming. We found that weighted prediction error (WPE) based dereverberation combined with generalized eigenvalue beamformer with power-spectral density (PSD) weighting masks generated by neural networks (NN) provides results approaching the clean close-microphone setup. Further improvement was obtained by re-training PLDA or the mask-generating NNs on simulated target data. The work shows that a speaker recognition system working robustly in the far-field scenario can be developed.

Index Terms— Speaker recognition, microphone array, beamforming, dereverberation, audio retransmission

1. INTRODUCTION

Performances of close-talk speaker recognition (SR) have significantly improved in the past 10 years, mainly due to the introduction of i-vectors [1]. However, far-field recognition still remains challenging. The reason is a distortion of the original speech signal. When a speaker talks in a room, sound waves propagate through air and get reflected on walls and obstacles. Owing to absorption of materials, they are attenuated and then they spread to the room again. It results in reverberation. Therefore, a microphone records multiple copies of the original speech.

Following [2], methods coping with reverberation can be divided into two groups: front-end- and back-end-based. As far as front-end-based approaches are considered, Cepstral Mean and Variance Normalization (CMVN) [3] of features is a straightforward option since it has been shown to cope well with convolutive distortion. However, a room impulse response (RIR) usually exceeds the length of a spectral analysis window, thus CMVN cannot tackle the effect of late reverberation. It can be then treated as an additive noise [4].

There have been other successful works related to reverberation-robust feature extraction. Zhang et al. [5] made use of deep neural networks (DNN). In this case, authors used DNN-based bottleneck features. The DNN is capable of transforming reverberant Mel-frequency cepstral coefficients (MFCC) to a new more discriminative space. They also proposed to map noisy features to their clean counterparts with denoising autoencoder (DAE).

When dealing with reverberation on a signal level, weighted prediction error (WPE) methods [6, 7] have proven to be very efficient at suppressing room acoustic effects. They are based on delayed linear prediction and are suitable for speech enhancement. Improvements in automatic speech recognition using the WPE are described for instance in [8].

Some methods (such as the WPE) may process both single- and multi-channel data. Therefore, multiple simultaneously recording microphones organized in microphone arrays [9] may be used when dealing with far-field recognition. The microphone arrays can serve as noise suppressors and at the same time means for dereverberation, as they mitigate the effects of reflected signals to some extent. Beamforming usually denotes steering the microphone arrays to a specific direction: among such techniques, the most intuitive one is delay-and-sum (DS) [10], using the fact that a sound wave impinges on different microphones at different time instants due to propagation delay. However, DS neglects the effect of room acoustics. Another beamformer is minimum variance distortionless response (MVDR), meant to suppress spatially correlated noise [9]. The MVDR beamformer is a result of optimization problem which minimizes the residual noise of the output subject to a distortionless constraint [11]. Recently, neural networks (NN) were incorporated into acoustic beamforming [12]. Heymann et al. employed them to estimate masks for noise and target signals that are used to compute power spectral density (PSD) matrices of noise and speech, respectively. Having them, the MVDR or generalized eigenvalue (GEV) beamformers [13] can be expressed.

The following text is structured as follows: In section 2, a new dataset is described. SR system parameters are given in section 3. Section 4 deals with performed experiments. Finally, conclusions are drawn in section 5.

The work was supported by Czech Ministry of Interior project No. VI20152020025 ”DRAPAK”, Grant Agency of the Czech Republic project No. GJ17-23870Y, and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project ”IT4Innovations excellence in science - LQ1602”.

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

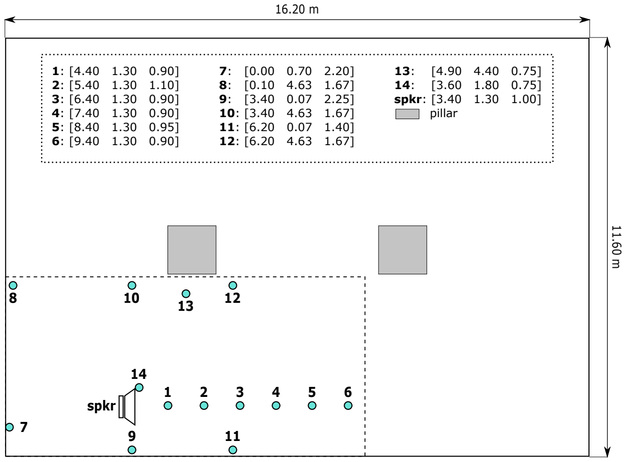

Fig. 1. *Floor plan of the room in which the retransmission took place. Coordinates are in meters and lower left corner is the origin. Dashed rectangle borders area displayed in Figure 3.*

2. TEST DATASET

To evaluate the impact of room acoustics on the accuracy of speaker recognition and efficiency of dereverberation methods, a proper dataset of reverberant audio is needed. An alternative that fills a qualitative gap between unsatisfying simulation (despite the improvement of realism [14]) and costly and demanding real speaker recording is retransmission. We can also advantageously use the fact that a *known* dataset can be retransmitted so that the performances are readily comparable with known benchmarks.

The retransmission took place in a room whose floor plan is displayed in Figure 1. The loudspeaker-microphone distance rises steadily for microphones 1…6 to study deterioration as a function of distance. Microphones 7…12 form a large microphone array to explore beamforming.

For this work, a subset of data released for NIST Year 2010 Speaker Recognition evaluations (SRE) was retransmitted. The dataset consists of 932 recordings with durations of three and eight minutes; 459 files include female voices and 473 include male voices. The total number of speakers is 300: 150 males and 150 females. Recordings from all microphones were synchronized at sample precision. The dataset is being gradually enlarged incorporating yet other rooms with different acoustics and recording procedures. BUT plans to release the dataset when finished; the version used to produce our results is available now on request.

3. SPEAKER RECOGNITION SYSTEM

In all the experiments we used an i-vector based speaker recognition system [1]. It comprises the classical components of feature extraction, universal background model represented by Gaussian mixture model (GMM-UBM), i-vector extraction, and probabilistic linear discriminant analysis (PLDA).

We used Mel-frequency cepstral coefficients (MFCC) of dimension 60 (including Δ and $\Delta\Delta$) as features. They were extracted from recordings in 10 ms steps (window length was 20 ms) and short time CMVN with 3-second window was implicitly applied to them. Such features were used for training of gender-independent GMM-UBM with 2048 components. The training dataset, which was a subset of PRISM set [15], consisted of 15600 telephone and microphone files including both female (1174) and male (813) speakers. Given a set of features and with the use of the GMM-UBM, sufficient statistics were computed. I-vectors, based on statistics, of dimension 600 were projected to 200-dimensional space using linear discriminant analysis (LDA). Latent variables in PLDA were of the same dimension. I-vector extractor and PLDA were trained on 86680 telephone and microphone files from PRISM set including 9663 female and 7013 male speakers.

4. EXPERIMENTS

All the results of experiments presented in this section are expressed in equal error rates (EER). For convenience, we show only female test data results. The baseline accuracy – 2.5% EER – was obtained on clean test data before the retransmission (original system, clean test data in Table 1).

4.1. Adverse effects of distance on speaker recognition

The aim of the first experiment was to discover whether there is a significant correlation between loudspeaker-microphone distance and SR accuracy. Therefore, we evaluated retransmitted test data captured by individual microphones with the original SR system. The results are displayed in Figure 2. All the microphones were intentionally divided into groups: *line*, *array* and *auxiliary*. Inter-microphone distance of sensors lying on *line* is one meter. All of them are in front of the loudspeaker and the line connecting them runs in the direction of sound wave propagation. Microphones seven to twelve form a microphone *array*. The remaining sensors are *auxiliary*. Regarding *line*, an approximate correlation deflected by local acoustic conditions is visible. The same holds for *auxiliary* microphones. The reason for lack of correlation in *array* is illustrated in Figure 3. Apparently, loudspeaker directivity pattern is the cause (see microphones 9 and 10 that are in line with the loudspeaker diaphragm).

4.2. System adaptation

Since the SR system consists of multiple components (section 3), adaptation may be performed on different stages of the processing chain. Our previous experiments revealed that mainly PLDA adaptation is of interest due to a great impact on results and low computational demands [16].

To adapt generatively trained PLDA, we performed training data augmentation by introducing close-to-target data to learn the far-field recordings channel. Since there is not much reverberant data for supervised PLDA training, we used image method simulation of room acoustics [17, 18] to obtain room impulse responses (RIR). The PLDA training data then

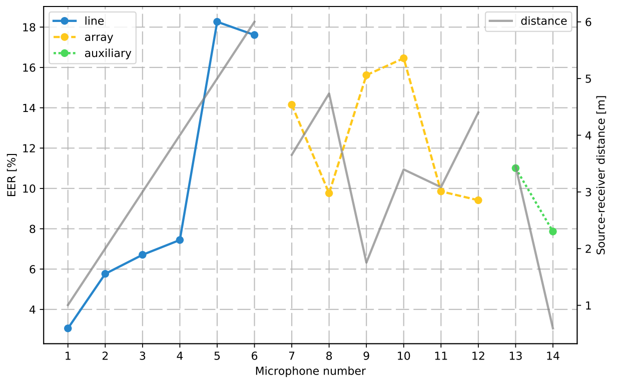

Fig. 2. *Correlation between loudspeaker-microphone distance and EER on female test data.*

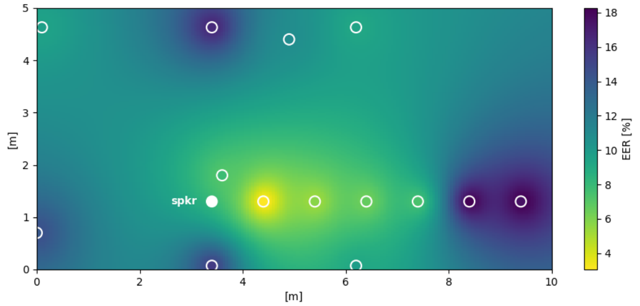

Fig. 3. *Floor plan cutout with interpolated EER values on female recordings. (The top-right corner values may be incorrect because we do not have enough data for interpolation.)*

consisted of (i) the original training data as described in section 3, (ii) a copy of the original training data (same number of files) convolved with RIRs of simulated rooms with random dimensions and random placement of microphones. Volumes of simulated rooms ranged from 18.4 m^3 to 600 m^3 (volume of the real room falls within this interval). The result of described adaptation is referred to as *adapt_simu* in Figure 4.

Next, we wanted to examine the adaptation using retransmitted data. Owing to the lack of such data we followed jackknifing schema: the test data were divided into two equally large parts from each microphone. Each of them contained the same number of both male and female speakers. Then the PLDA was trained on the original data with the first part of the test data (the original training dataset was extended by 6524 files) and then tested on the second part of the test data. This was repeated with swapped splits and the outcomes averaged. The results are shown in Figure 4 – *adapt_retrans*. It is visible that the performance is worse compared to *adapt_simu*. However, it is worth mentioning that relative average improvement of EER for *adapt_simu* is 40.3% and 32.5% for *adapt_retrans*. However, *adapt_simu* PLDA saw much more reverberant data than *adapt_retrans* which might be the reason for having bigger improvement. We created a concatenated condition with both simulated and retransmitted data which is denoted as *adapt_both* and we see that there is a nice improvement of the *adapt_simu* which shows that the in-domain data helps. It should be also noted that *adapt_retrans* assumes knowledge of the target room and positions of microphones; none of them might be known in a real scenario.

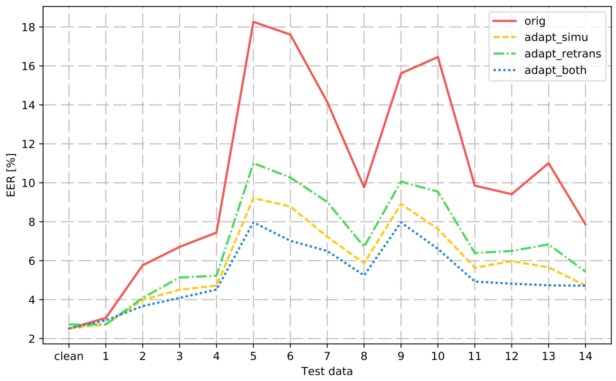

Fig. 4. *Comparison of system adaptation methods in terms of EER. Only female test recordings are considered.*

4.3. Dereverberation

Two techniques for dereverberation were explored: weighted prediction error (WPE) and denoising/dereverberation neural network autoencoder (DNS). For application of WPE, we used Matlab p-code[1] by the authors of [6, 7].

The autoencoder used for denoising/dereverberation consists of three hidden layers with 1500 neurons in each layer. The input of the autoencoder was a central frame of a log-magnitude spectrum with a context of +/- 15 frames (in total 3999-dimensional input). The output is a 129-dimensional enhanced central frame. We used Mean Square Error (MSE) as objective function during training. Fisher English database parts 1 and 2 were used for training the autoencoder, approximately 1800 hours of audio. The datasets were artificially corrupted with noise on SNR level 0-21dB from Freesound library [2] and RIRs were taken from AIR database [19].

Results obtained using the original PLDA (no adaptation) to capture only the effect of signal pre-processing are shown in Figure 5. It can be seen that WPE (*wpe10*) achieved great suppression of late reverberation, especially for close-to-source microphones. However, when reverberation time prolonged, WPE even caused accuracy deterioration. The filter of *wpe10* had 10 coefficients. To deal better with long reverberations, we extended the number to 15 (*wpe15*). It improved all the results, not only those that suffered degradation. On the contrary, the neural network denoising (*dns*) achieved very stable improvements.

4.4. Beamforming and combination with dereverberation

In this section, effects of beamforming and dereverberation applied to microphones 7 to 12 are presented. In Table 1, we

[1] `http://www.kecl.ntt.co.jp/icl/signal/wpe`

[2] `http://www.freesound.org`

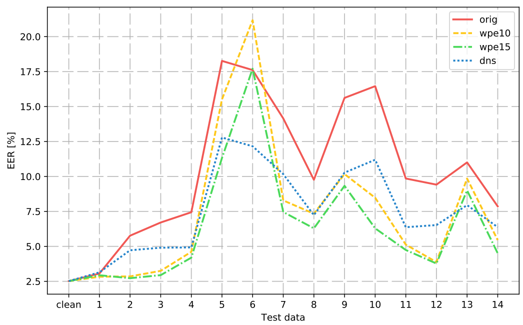

Fig. 5. *Comparison of dereverberation methods in terms of EER. Only female test recordings are considered.*

show all the results and we also compare different systems: the original, the system retrained with simulated data (section 4.2), the system adapted with dereverberated data. The only difference between training data for two last systems is that for the latter, reverberant data were processed by corresponding dereverberation method to tackle acoustic channel.

A basic delay-and-sum (*DS*) uses generalized cross-correlation with phase transform weighting (GCC-PHAT) in order to estimate time difference of arrival (TDOA) as it was shown to be less prone to effects of reverberation [20]. Minimum variance distortionless response beamformer (*MVDR*) assumes noise to be diffuse [21] rather than directional as there was no point source of noise during retransmission. We also tested BeamformIt tool [22] which performs weighted delay-and-sum and other advantageous signal processing. We found the following techniques useful: reference microphone computation, channel weighting, Viterbi decoding and N-best GCC-PHAT values consideration. All of them are referred to as *BeamformIt*. From the results shown in the middle part of Table 1, it can be seen that none of these methods was able to outperform the best individual microphone.

FW_GEV refers to the generalized eigenvalue beamformer that uses PSD masks estimated by a feed-forward neural network. First, we used the NN[3] trained by the authors of [12]. Despite being trained mainly to cope with noise, the beamformer was able to deliver promising results on our reverberant test data. To tackle reverberation, we altered training data and re-trained the NN (*FW_GEV_rever*). The ideal speech masks were computed out of the clean data convolved with the first 50 ms of random RIRs (this was shown to be beneficial in [23]). Noise masks were computed analogically taking the rest of RIRs into account. *FW_GEV_rever* brought a substantial improvement especially when no dereverberation technique was used. Overall, the best results were obtained with the combination of WPE (15 coefficients) and *FW_GEV_rever* (only 4.2% EER relatively worse than in the clean data case; the best single microphone results on reverberant data was 274.2% relatively worse for comparison).

Table 1. *Beamforming and dereverberation methods and their combinations. The EER values in percent were obtained by evaluating female test recordings. "Best" and "worse" denote the results from the best and worst performing individual microphones 7 to 12. WPE refers to the 15-coefficient WPE, DNS to the NN denoising/dereverberation.*

Test data		Original system	Simulated data adapt.	Dereverb. data adapt.
clean		2.5158	2.5158	-
reverberant	best	9.4150	5.6383	-
	worse	16.4570	8.9097	-
DNS	best	6.3707	5.0314	4.0881
	worse	11.1948	8.2789	7.4468
WPE	best	3.8783	3.6689	3.5639
	worse	10.1678	9.2244	7.8674
DS		14.1456	9.0147	-
MVDR		13.6241	7.4422	-
BeamformIt		9.4339	6.0797	-
FW_GEV		10.0716	5.5556	-
FW_GEV_rever		**7.5408**	**4.931**	-
DNS + DS		9.3292	6.7086	6.1845
DNS + MVDR		9.4456	6.4989	5.7545
DNS + BeamformIt		8.4906	6.8365	6.1845
DNS + FW_GEV		7.3592	5.6603	5.2411
DNS + FW_GEV_rever		**6.2863**	**4.2978**	**4.5041**
WPE + DS		6.1845	6.0797	5.6577
WPE + MVDR		6.1799	5.0314	4.9267
WPE + BeamformIt		5.0297	4.3047	4.0881
WPE + FW_GEV		2.8276	**2.7253**	**2.6206**
WPE + FW_GEV_rever		**2.7253**	2.8303	2.7253

5. CONCLUSIONS

In this work, we explored multiple beamforming and dereverberation techniques along with system adaptation to deal with a far-field speaker recognition. Moreover, we introduced a new dataset of recordings retransmitted in real-world acoustic conditions.

We have shown that combinations of the discussed methods can deliver significant improvements. The best results were obtained by applying WPE dereverberation and subsequent neural network based GEV beamforming while using WPE data adapted PLDA. The EER was then only 4.2% relatively worse than the EER measured on clean data.

Only one room was considered in the experiments. Therefore, applicability in different acoustic conditions should be further studied as well as realistic (not re-recorded) data. Another challenge will be non-synchronous recordings and moving speakers.

[3] `https://github.com/fgnt/nn-gev`

6. REFERENCES

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011, ISSN: 15587916.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[3] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[4] Q. Jin, T. Schultz, and A. Waibel, "Far-Field Speaker Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2023–2032, 2007.

[5] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, and M. Iwahashi, "Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, 2015.

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[7] T. Yoshioka and T. Nakatani, "Generalization of Multi-Channel Linear Prediction Methods for Blind MIMO Impulse Response Shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[8] T. Yoshioka and M. J. F. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech & Language*, vol. 31, no. 1, pp. 65–86, 2015.

[9] K. Kumatani, J. McDonough, and B. Raj, "Microphone Array Processing for Distant Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012, ISSN: 10535888.

[10] I. McCowan, "Microphone Arrays : A Tutorial," 2001.

[11] M. Souden, J. Benesty, and S. Affes, "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010, ISSN: 15587916.

[12] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 196–200, IEEE.

[13] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[14] M. Ravanelli, P. Svaizer, and M. Omologo, "Realistic Multi-Microphone Data Simulation for Distant Speech Recognition," 2016, pp. 2786–2790.

[15] L. Ferrer, H. Bratt, L. Burget, J. Černocký, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matějka, O. Plchot, et al., "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," *https://code.google.com/p/prism-set/*, 2012.

[16] O. Glembek, J. Ma, P. Matějka, B. Zhang, O. Plchot, L. Burget, and S. Matsoukas, "Domain Adaptation Via Within-class Covariance Correction in I-Vector Based Speaker Recognition Systems," in *Proceedings of ICASSP 2014*, 2014, pp. 4060–4064.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979, ISSN: 0001-4966.

[18] E. A. P. Habets, "Room Impulse Response Generator," September 2010.

[19] "Aachen impulse response database," http://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen-impulse-response-database/.

[20] J. Chen, J. Benesty, and Y. (Arden) Huang, "Time Delay Estimation in Room Acoustic Environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–20, 2006, ISSN: 16876172.

[21] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New Insights Into the MVDR Beamformer in Room Acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2010.

[22] X. Anguera, C. Wooters, and J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[23] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, vol. 46, pp. 374–385, 2017.