Interspeech 2018
2-6 September 2018, Hyderabad

Dereverberation and Beamforming in Robust Far-Field Speaker Recognition

Ladislav Mošner, Oldřich Plchot, Pavel Matějka, Ondřej Novotný and Jan "Honza" Černocký

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

{imosner, iplchot, matejkap, inovoton, cernocky}@fit.vutbr.cz

Abstract

This paper deals with robust speaker verification (SV) in far-field sensing. The robustness is verified on a subset of NIST SRE 2010 corpus retransmitted in multiple real rooms of different acoustics and captured with multiple microphones. We experimented with various data preprocessing steps including different approaches to dereverberation and beamforming applied to ad-hoc microphone arrays. We found that significant improvements in accuracy can be achieved with neural network based generalized eigenvalue beamformer preceded by weighted prediction error dereverberation. We also explored the effect of data augmentation by adding various real or simulated room acoustic properties to the Probabilistic Linear Discriminant Analysis (PLDA) training dataset. As a result, we developed a speaker recognition system whose performance is stable across different room acoustic conditions. It yields 41.4% relative improvement in performance over the system without multi-channel processing tested on the cleanest microphone data. With the best combination of data preprocessing and augmentation, we obtained a performance close to the one we achieved with the original clean test data.

Index Terms: speaker verification, beamforming, dereverberation, autoencoder

1. Introduction

Robustness of speaker verification (SV) systems is a crucial property especially when they are used in real conditions. Even though the reliability of recognition while recording with far-field microphones still remains challenging, a limited number of works have been published [1, 2]. When remote microphones are employed to capture speech signals, they inevitably record attenuated copies of the original signal that are present in enclosures as a result of the multipath propagation of a source signal. In other words, the channel between source and receiver corresponds to a system described by room impulse response (RIR). Therefore, the channel differs for every position even in the same room. RIR is usually much longer than the analysis window which prevents methods like cepstral mean and variance normalization (CMVN) applied to features from dealing well with late reverberation [1].

Restoration of a source signal is an objective of the first type of approaches that cope with effects of room acoustics. Various speech enhancement methods were developed [3] over years. In this work, we will focus on two dereverberation techniques differing in their nature. The first one is weighted prediction error (WPE) based on delayed linear prediction [4, 5]. It is capable of processing both single- and multi-channel data. The denoising/dereverberation autoencoder [6] will be the second method of our interest.

Since our dataset consists of multichannel recordings from irregularly shaped microphone arrays, beamforming techniques may be used. Microphone arrays are spatial filters capable of enhancing the desired signal and suppressing interfering sounds at the same time, which makes them also means for dereverberation. Usually, microphones in arrays are positioned in a regular topology. However, ad hoc microphone arrays, where no prior information about microphone positions is available, are more realistic. They must work blindly to locate the direction of interest. Various approaches were developed to tackle this problem [7]. We will use estimation based on generalized cross-correlation with phase transform (GCC-PHAT) in combination with a basic delay-and-sum (DS) beamformer. Recently, usage of neural networks in acoustic beamforming emerged [8, 9]. We will make use of the neural generalized eigenvalue beamformer by Heymann et al. [9].

The second approach focuses on a backend. In our SV system, Probabilistic Linear Discriminant Analysis (PLDA) is used as the backend and in order to adapt it to a new channel, introduced by either room conditions or preprocessing, we designed multiple training datasets to capture these channel effects and compared their impact on system adaptation.

This work is a continuation of our initial experiments in [10] where we analyzed aforementioned methods on limited amount of data. In this work, we focus on system robustness and consistency across multiple environments as we perform tests on data obtained from three different rooms. On top of the analysis presented in [10], we made use of multichannel real and simulated data also in the PLDA training further improving the adaptation to the test scenario. We also examined the correlation of the SV system accuracy with D50 acoustic parameter and explored a possible scenario for SV when we select the best possible single channel and avoid the multi-channel processing.

2. Speaker recognition system

We used a standard speaker verification system based on Mel-frequency cepstral coefficients (MFCCs) and i-vectors [11]. 19 MFCC coefficients were extracted along with log-energy from 25 ms windows with 15 ms overlaps. MFCCs were augmented with their Δ and $\Delta\Delta$ coefficients providing 60-dimensional feature vectors that were further subjected to the cepstral mean and variance normalization (CMVN) with 3-second sliding window. A gender-independent 2048 component GMM-UBM was trained on a subset of PRISM dataset [12] which consists of 15600 telephone and microphone files including both female (1174) and male (813) speakers. I-vector extractor was trained on 86629 telephone and microphone files from PRISM set including 9663 female and 7013 male speakers. Dimensionality of i-vectors was set to 600 and the i-vectors were later pro-

The work was supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK", Google Faculty Research Award program, Czech Science Foundation under 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".

10.21437/Interspeech.2018-2306

jected to 200-dimensional space by means of Linear Discriminant Analysis (LDA). PLDA classifier was trained on the same dataset as the i-vector extractor.

3. Experimental setup

Since this work aims at exploration of speaker verification in far-field scenarios, the subset of data released for NIST Year 2010 Speaker Recognition evaluations (SRE) was retransmitted in real rooms. The retransmitted dataset consists of 459 three or five minutes long recordings spanning 150 female voices. 302 utterances were originally recorded with microphones and 157 over the telephone channel.

During the retransmission, recordings were simultaneously captured by 8 or 6 microphones placed in various locations of a room. We consider these microphones as an ad hoc microphone array without any particular shape. The retransmission was performed in three different rooms with reverberation times (T30) 0.75 s, 0.53 s and 0.65 s. A placement of loudspeaker and microphones varied across rooms.

3.1. Denoising/dereverberation autoencoder and WPE

The DNN autoencoder used for dereverberation maps frames of a log-magnitude spectrum to their dereverberated counterparts. The exact description of the architecture is given in [6]. In this text, the network will be referred to as *DNS*.

In order to perform weighted prediction error (*WPE*), we divided audio signals into frames of 32 ms by shifting the Hann window by 8 ms. The number of filter coefficients was set to 15.

3.2. Beamforming techniques

In order to perform beamforming, we made use of two different methods. The first one is a basic delay-and-sum (*DS*) beamformer [13], the second one is a generalized eigenvalue (GEV) beamformer [14] which uses a neural network (NN) to predict masks applied to spectra when computing power spectral density matrices (PSD) [9].

For the delay-and-sum processing, the input signal is divided into 500 ms overlapping segments multiplied by Hann window. The time difference of arrival (TDOA) is determined for every frame by a position of the maximum of a generalized cross-correlation with phase transform (GCC-PHAT) output.

For the GEV beamformer mask estimation, we used a feed-forward NN whose architecture is defined in [9]. We explored two training schemes. In both of them, we used 7138 recordings of the simulated training data defined by the 3rd CHiME challenge [15]. In the first scheme, the data were prepared[1] according to [9]. The beamformer using this NN will be referred to as *FW_GEV*. The processing was designed mainly to tackle noise. We prepared also another dataset aimed more at reverberation. It was obtained by convolving the original training recordings with parts of random room impulse responses (RIR). Following [16], the result of convolution with the first 50 ms of RIR was considered to be a clean audio. Convolution with the rest resulted in a signal used for "noise" mask computation. The beamformer will be denoted as *FW_GEV_rever*.

3.3. PLDA training data augmentation

In our previous experiments [17], we observed that in order to adapt SV system, the PLDA is a key part to focus on. Therefore, we re-trained it on different datasets of specific properties. In order to focus on a channel variability, all datasets contain the same set of 16676 speakers.

orig The original dataset described in section 2. We consider the resulting system to be our baseline.

simu Dataset designed to be closer to the test data in terms of an acoustic channel. It contains the same recordings (86629) as the *orig* dataset. However, all of them were convolved with simulated room impulse responses. We used the image method [18, 19] to obtain RIRs. Room dimensions were chosen randomly for every recording. Minimum room size was set to $2.0 \times 4.0 \times 2.3$ m, maximum to $10.0 \times 12.0 \times 5.0$ m. The placement of the source and the receiver was random as well.

orig + multich We selected data from previous NIST SREs (2005, 2006, 2008, 2010) that were simultaneously recorded over multiple (4 to 14) microphones. In total, we selected 3490 utterances from 448 speakers already present in the *orig* dataset. The recordings were processed exactly the same way as the test data (see section 4). Resulting single-channel recordings were then combined with the *orig* ones. Overall, this set contains 90119 recordings.

simu + multich This dataset is almost the same as *orig + multich*. The only difference is that the *simu* data are used instead of *orig*.

derever adapt The base for *derever adapt* is *orig + simu*. The simulated part was further processed by appropriate dereverberation method.

4. Results

Results of our experiments are listed in Table 1 in terms of equal error rate (EER) computed on pooled scores from all three rooms. The columns correspond to systems (training datasets). The table is further divided into row sections according to the test data preprocessing. Data (*clean*) for our baseline system are later used for retransmission. The baseline EER of 2.71% was obtained by evaluation of the clean test data with the *orig* system. The aim is to approach such accuracy on reverberant data. We did not apply any test data preprocessing for systems in the rows denoted as *reverberant*. In those rows, a substantial deterioration of accuracy can be seen. For convenience, only the best and the worst performing single microphones are shown.

4.1. Reverberant test data and dereverberation

It was shown that the definition (D50) acoustic parameter [20] closely correlates with speaker recognition performance [21]. In Figure 1, the EER values are displayed as a function of D50 for all microphones in all rooms. The baseline SV system was used to obtain EERs. It can be seen that the performances vary substantially across rooms and the correlation with D50 approximately holds. Knowing the D50 values beforehand can be beneficial for selection of the best microphone and use of a simple single-channel speaker verification. The accuracy obtained with the microphones picked according to the best and the worst (the highest and the lowest) D50 values are shown in brackets in the row with results on reverberant data in Table 1 (section *basic*). The EERs chosen by taking global minimum and maximum over all microphones are also in the table (values without brackets, section *basic*). They confirm that the choice according to D50 is reasonable.

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

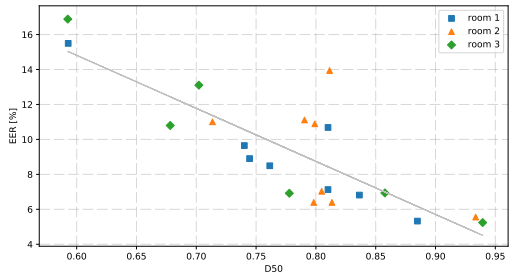

Figure 1: *Distribution of EERs with respect to the D50 values corresponding to individual microphones in tested rooms.*

In the *dereverb* section, it can be seen that when testing on dereverberated data from the best microphone, an improvement is achieved. When using WPE, the difference in EER is greater. It again holds that the best performing microphone (according to accuracy) corresponds to that selected by the D50 value. In the case of WPE, the best resulting accuracy (2.62% EER) was even better than for the original test data (2.71% EER). It is, however, worth mentioning that the knowledge of D50 values is idealistic. They depend on positions of both speaker and microphone since they are computed from RIRs. Therefore, the placement must be known before testing. Moreover, the acoustic conditions of the target room may not be known at all. Regarding the worst performing microphones, they do not need to correlate with the lowest D50 as shown in the table. We observed that for certain acoustic conditions, dereverberation may even cause accuracy deterioration (WPE). Therefore, a different microphone can become the worst. In our experiments, the NN based dereverberation yielded more stable results.

4.2. Preprocessing with beamforming

We compare two approaches that are rather different in complexity but may work blindly, requiring neither speaker nor microphone locations. The results obtained on beamformed test data with differently adapted SV systems are shown in Figure 2 and in the *beamf.* section of Table 1.

Division of the systems into two clusters is visible. The original system can be improved by augmenting the training data by applying the same pre-processing that is applied to the test data. However, it seems to be more important to provide the PLDA with recordings convolved with RIRs during training. The possible explanation is that beamforming performs a combination of linearly filtered reverberant data. The channel of such data could be closer to the linearly filtered training data (using RIRs) than the channel of the original data. Further improvement can be achieved by the introduction of beamformed real multichannel data to reverberant dataset.

4.3. Combination of beamforming and dereverberation

In this part, we explore the effects of combinations of beamforming and dereverberation methods. According to our experiments, the order in which the methods are applied is important.

Regarding the denoising/dereverberation autoencoder, we observed that neural network processing should be preceded by beamforming. In case the DS is applied after NN processing, it fails on individually dereverberated signals that presumably lost some phase information. Mask estimation in FW_GEV also performs poorly as it is trained on data from a different domain and it should estimate which frequency bins are dominated by

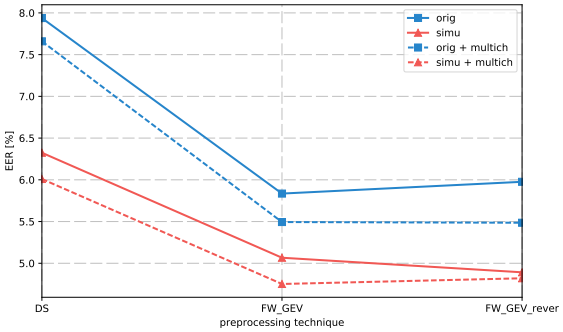

Figure 2: *Performance of SV systems when beamforming is applied to the test data.*

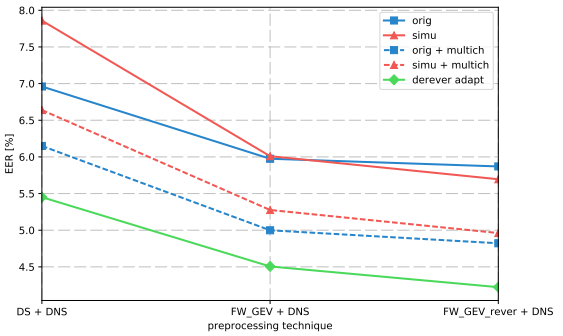

Figure 3: *Performance of SV systems when beamforming and subsequent autoencoder based dereverberation is applied to the test data.*

noise/reverberation after denoising/dereverberation.

In Figure 3 and in the *DNS + beamf.* section of Table 1, it can be seen that evaluating with the *simu* system is not sufficient as it does not bring overall improvement. The acoustic channels present in the *orig* and *simu* data seem to be different from that introduced by non-linear NN processing. Indeed, introduction of the training data processed by beamforming and subsequent DNS brings improvements (dashed lines). Since *orig + multichan* works promisingly, we trained another system making use of simulation to produce more training data that are subsequently dereverberated by DNS (*derever adapt*). Further average absolute improvement of 0.6% EER was then achieved.

We also experimented with WPE instead of DNS. In this case, WPE should precede beamforming. Since WPE is able to process multichannel signals and output again multichannel signals, it is perfectly suitable for this scheme[2].

The *simu* system yields by far the worst performance among systems shown in Figure 4 (see also the *WPE + beamf.* seciton of Table 1). It corresponds to the ability of WPE to dereverberate signals. In the previous experiments, we observed that the addition of multichannel real data processed by means of dereverberation in combination with beamforming to the training data can bring improvement in performance. When such data, originating from real conditions, were added to *simu*, creating *simu + multichan* in this experiment, considerable improvement in EER was observed. However, the change was not sufficient and the *orig* system still worked better. As expected, the performance of the *orig* system augmented with multichannel data processed consistently with the test data (*orig + multichan*) improved over the *orig* system performance. While using

[2]We also observed that multichannel WPE is more efficient.

Table 1: *Results of the experiments in terms of EER [%]. Values in brackets were selected according the best/worst values of D50.*

	Test data		orig	simu	orig + multich	simu + multich	derever adapt
basic	clean		2.71	2.73	-	-	-
	reverberant	best	5.24 (5.24)	4.31 (4.59)	-	-	-
		worst	16.88 (16.88)	12.37 (12.37)	-	-	-
dereverb.	DNS	best	3.90 (3.90)	4.30 (4.30)	-	-	3.64 (3.64)
		worst	12.57 (10.17)	12.66 (10.82)	-	-	9.12 (7.97)
	WPE	best	2.73 (2.73)	3.67 (3.67)	-	-	2.62 (2.93)
		worst	19.71 (9.12)	18.34 (8.60)	-	-	17.49 (8.48)
beamf.	DS		7.94	6.32	7.66	**6.01**	-
	FW_GEV		5.84	5.07	5.49	**4.75**	-
	FW_GEV_rever		5.98	4.89	5.49	**4.82**	-
DNS + beamf.	DS + DNS		6.96	7.86	6.15	6.64	**5.45**
	FW_GEV + DNS		5.97	6.01	5.00	5.28	**4.51**
	FW_GEV_rever + DNS		5.87	5.70	4.82	4.96	**4.22**
WPE + beamf.	WPE + DS		5.10	5.91	4.89	5.14	**4.68**
	WPE + FW_GEV		3.32	3.74	3.21	3.59	**3.07**
	WPE + FW_GEV_rever		3.35	3.95	3.18	3.39	**3.15**
combined	WPE + DS + DNS		5.74	7.79	**5.00**	6.01	-
	WPE + FW_GEV + DNS		3.63	4.47	**3.46**	3.74	-
	WPE + FW_GEV_rever + DNS		3.66	4.37	**3.11**	3.60	-

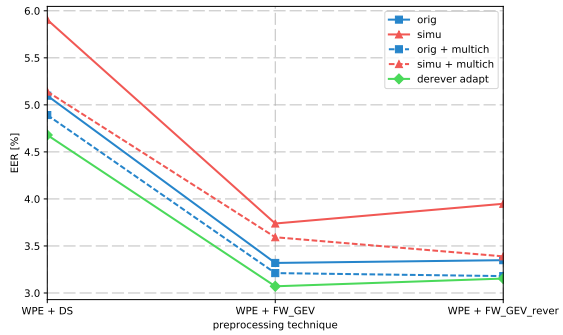

Figure 4: *Performance of SV systems when WPE dereverberation and subsequent beamforming is applied to the test data.*

DNS denoising/dereverberation, *derever adapt* improved upon *orig + multich* by substantial margin. However, when the WPE is applied instead, the difference between those two systems is almost negligible. It led to an important outcome of the experiments: acquisition of sufficient amount of multichannel labeled data is not straightforward (it was demonstrated by *orig + multich*). However, one can make use of simulated data (despite their realism is limited by simplifications of the image method) that can be easily generated in great amounts and obtain a system with comparable performance to the one using real data.

4.4. Overall results

Intuitively, application of all preprocessing steps is an option as well. According to the ordering of methods determined in section 4.3, we first applied WPE, then beamforming and finally DNS. However, it can be seen in the *combined* block of Table 1, that by application of such preprocessing, no further improvement over *WPE + beamf.* preprocessing is achieved.

Overall, the best results are achieved when *WPE + beamf.* is used for the test data preprocessing and either *orig + multich* or *derever adapt* systems are used for verification (see Table 1). Out of the beamforming techniques, FW_GEV is the best in this setting. Training the GEV NN to tackle reverberation does not appear to bring any improvement probably due to the dereverberation capability of preceding WPE. In terms of EER, the best combination yields 3.07% EER, which is 13.5% relatively worse result compared to the accuracy on the clean test data. However, it also means 41.4% relative improvement over the best possible microphone.

5. Conclusions

We have analyzed multiple approaches to deal with the far-field speaker verification. Our aim was to develop a robust system that can work with consistent performance in various acoustic conditions. In the data preprocessing phase, we explored effects of the WPE and autoencoder dereverberation. Moreover, we made use of multichannel data from ad hoc microphone arrays and we applied blind beamforming (delay-and-sum and generalized eigenvalue beamformers). In order to adapt the SV system, we used both simulated and real reverberant data.

We observed that applying the WPE and then choosing an appropriate microphone in tests can deliver approximately the same accuracy as the baseline. However, relying on choosing the best microphone is questionable as a prior knowledge about the settings of the actual room is rarely available and the existence of such a good microphone is not guaranteed. On the other hand, promising results can be achieved when the WPE succeeded by the GEV beamformer are applied to the microphone array data. In this case, systems adapted with a small portion of real data or a large amount of simulated data both dereverberated with WPE perform almost equally.

In the future, the behavior of the WPE dereverberation should be studied as we observed degradation of results for some microphones. The D50 acoustic parameter could also be used for selection of microphones that form a microphone array.

6. References

[1] 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. [Online]. Available: http://ieeexplore.ieee.org/document/4291604/

[2] M. K. Nandwana, J. van Hout, M. McLaren, A. Stauffer, C. Richey, A. Lawson, and M. Graciarena, "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings," in *Proc. Interspeech 2018*, September 2018.

[3] D. S. Kulkarni, R. R. Deshmukh, and P. P. Shrishrimal, "A Review of Speech Signal Enhancement Techniques," *International Journal of Computer Applications*, vol. 139, no. 14, pp. 23–26, 2016. [Online]. Available: http://www.ijcaonline.org/research/volume139/number14/kulkarni-2016-ijca-909507.pdf

[4] 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. [Online]. Available: http://ieeexplore.ieee.org/document/5547558/

[5] 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. [Online]. Available: http://ieeexplore.ieee.org/document/6255769/

[6] O. Plchot, L. Burget, H. Aronowitz, and P. Matějka, "Audio enhancing with DNN autoencoder for speaker recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5090–5094.

[7] I. Himawan, I. McCowan, and S. Sridharan, "Clustered Blind Beamforming From Ad-Hoc Microphone Arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661–676, May 2011.

[8] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5745–5749.

[9] 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)*. IEEE, 2016, pp. 196–200. [Online]. Available: http://ieeexplore.ieee.org/document/7471664/

[10] L. Mošner, P. Matějka, O. Novotný, and J. Černocký, "Dereverberation and Beamforming in Far-Field Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[11] 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. [Online]. Available: http://ieeexplore.ieee.org/document/5545402/

[12] 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.

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

[14] 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. [Online]. Available: http://ieeexplore.ieee.org/document/4244540/

[15] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 2015, pp. 504–511.

[16] 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. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0885230816300730

[17] 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 *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4060–4064.

[18] J. Allen and D. 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.

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

[20] H. Kuttruff, *Room acoustics*, 5th ed., 2009, ISBN: 978-0415480215.

[21] I. Peer, B. Rafaely, and Y. Zigel, "Room Acoustics Parameters Affecting Speaker Recognition Degradation Under Reverberation," in *2008 Hands-Free Speech Communication and Microphone Arrays*, May 2008, pp. 136–139.