BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2017
http://excel.fit.vutbr.cz

Microphone Arrays for Speaker Recognition

Ladislav Mošner*

Abstract
This paper addresses the problem of remote speaker recognition (SRE). Nowadays, systems working with signals from close-talking microphones yield very good results. However, in the case of remotely obtained data, their accuracy decreases considerably. Therefore, we explore the applicability of microphone arrays for recovery from errors of SRE system introduced by the room reverberation, where microphone arrays are purposely positioned set of microphones. We discuss two approaches which are both based on microphone arrays – beamforming (delay-and-sum) for enhancement of the input signals and the retraining of the SRE system components with the beamformed data. By the combination of both techniques, we have achieved significant improvement of accuracy in comparison with the results obtained with the single-microphone recordings.

Keywords: Speaker recognition — Microphone arrays — Beamforming — Room impulse response

Supplementary Material: *N/A*

*xmosne01@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Nowadays, the systems for speaker recognition (SRE) achieve accuracies that merit the attention. It is mainly due to a great effort that was put into the research of such systems. As the result, the mathematically sophisticated but usable systems that are based on the i-vector representation of recordings came into existence. However, these systems mostly require signals acquired from the close-talk microphones, which restrains systems from being used in far-field scenarios. In such cases, the accuracy decreases substantially, which is caused by disturbing with room noise and reverberation.

Solving the outlined problem is the motivation for this work. First, we will explore how much the SRE accuracy deteriorates with the remote microphones. Then, the possibilities for improvement will be discussed. We will make use of the microphone arrays that allow for the enhancement of the signal coming from a particular direction by beamforming. Multiple beamforming methods exist. The one that will be used in this work is the *delay-and-sum* method. Another approach to the accuracy improvement is the adaptation of a system to new conditions.

We will show that the combination of both approaches is beneficial. Even though we did not achieve the same accuracy as with the close-talk scenario, the improvement of the far-field scenario was significant.

To give a basic theoretical background, section 2 covers essential principles of current SRE systems and possibilities of microphone arrays along with the description of beamforming. In section 3, we explain the utilized dataset and need for data simulation. In section 4, beamforming preprocessing and modifications that led to accuracy improvements will be summarized. Finally, section 5 presents the experiments.

2. Background

2.1 I-vector Based Speaker Recognition

In this work, we will use the current state-of-the-art approach to speaker recognition [1]. It is well established and has been used with some modifications for a few years and it still yields supreme results. The whole system can be seen as a chain consisting of major "blocks", namely *feature extraction*, *Gaussian mixture universal background model* (UBM), *i-vector extraction* and *probabilistic linear discriminant analysis*, which produces final scores. The presented sequence is depicted in Figure 1. A brief description of the blocks follows:

Feature extraction As it is inconvenient to work directly with the raw audio signal, there is a need for conversion to a suitable float-vector representation with lowered dimension. Therefore, the feature extraction methods are designed to reduce the dimensionality and produce the feature vectors, typically 1 per 10 ms step. In speech processing, Mel-Frequency Cepstral Coefficients (MFCC) are very common features [2].

Gaussian mixture modeling The feature vectors acquired in the previous step are modeled with the *Gaussian mixture model* (GMM). GMM is a *generative probabilistic model* [3, 4], which consists of C weighted normal probability distribution functions (PDF) and is described by three parameters – $\boldsymbol{w}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, where $\boldsymbol{w}$ is a vector with component weights, $\boldsymbol{\mu}$ denotes a supervector obtained by concatenation of per-component mean vectors $\boldsymbol{\mu}_c$, where $c = 1,\ldots,C$. The last parameter $\boldsymbol{\Sigma}$ is a block diagonal matrix, in which the diagonal consists of the covariance matrices $\boldsymbol{\Sigma}_c$ for each component.
Typically, the training dataset consists of utterances from many different speakers. The resulting GMM hence covers the space of all the training speakers and is called *universal background model* (UBM).

I-vector extraction For the purposes of speaker recognition, a recording is represented by one vector with a fixed dimensionality regardless of the duration of the recording – i-vectors [1, 5]. The relation between the mean supervector m, that we would obtain by maximum a posteriory (MAP) adaptation of UBM to an utterance, and the i-vector $\boldsymbol{\phi}$ is given by [3]

$$\boldsymbol{m} = \boldsymbol{\mu} + \boldsymbol{T}\boldsymbol{\phi}, \tag{1}$$

where $\boldsymbol{\mu}$ is a supervector of UBM means and $\boldsymbol{T}$[1] is a matrix which defines a subspace (*total variability space*) of directions in which the means get adapted. With the UBM we extract the *sufficient statistics* from recordings. These are used in $\boldsymbol{T}$ matrix training and i-vector extraction.

Probabilistic linear discriminant analysis (PLDA) PLDA was proposed by Prince and Elder [6] for the task of face recognition. Then it was successfully adopted in a field of speaker recognition [7]. It is used for classification of i-vectors. Because the i-vectors include information about a speaker, but also about gender, recording conditions (channel), etc., the PLDA model comprises matrices spanning the speaker and channel subspaces [8, 9]. These matrices are utilized in the phase of scoring, which is very fast and symmetric. PLDA then produces a score, which represents the similarity of speaker voices in two recordings.

The system that was presented is available in the BUT Speech@FIT group[2]. We implemented own i-vector extractor and PLDA training scripts. The script for PLDA is used in this work. However, due to a slower convergence of i-vector extractor training, we finally did not use our version.

2.2 Microphone Arrays and Beamforming

When it comes to a processing of speech recorded by distant microphones, the use of a single microphone is inconvenient, as it records all the noises and other unwanted speech signals coming from different directions. To handle the problem, microphone arrays are a good choice. The microphone array is a set of microphones typically organized in a topology. Often they form a circle or a grid in the two-dimensional plane.

A microphone array can be interpreted as a *spatial filter*, which is capable of enhancing the signal coming from a specific direction, while it attenuates the noise and competing speech coming from other directions [10]. The filter can be described by a *directivity pattern* [11] or a *beampattern* [10], which specifies the array response as the function of frequency and direction of arrival. For a uniform linear array[3] (ULA) and specific frequency, the directivity pattern is depicted in Figure 2. The lobe around the maximum is called *main lobe*, other lobes are *sidelobes* [11].

[1] $\boldsymbol{T}$ is sometimes referred to as an i-vector extractor [3].

[2] http://voicebiometry.org/

[3] Uniform linear array stands for a microphone array that consists of microphones positioned on a line. The spacing between neighboring microphones is uniform.

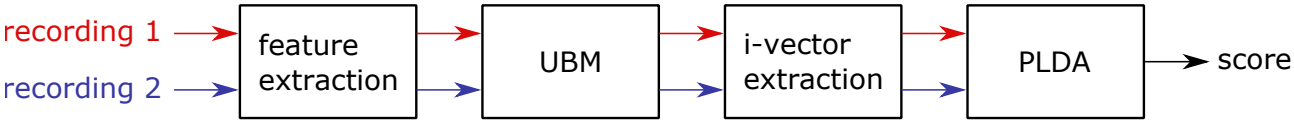

Figure 1. A block diagram of used speaker recognition system.

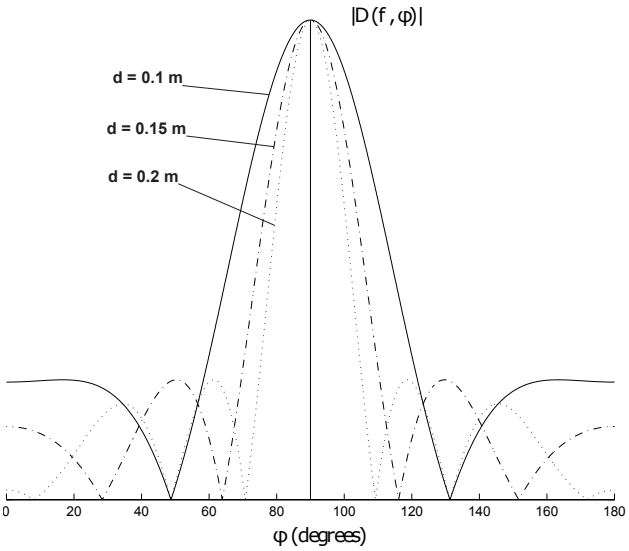

Figure 2. Directivity patterns for a uniform linear array. The spacing d between microphones affects the shape. Taken from [11].

The ability of microphone arrays which we are interested in is called *beamforming*. This technique implements the *shaping* and *steering* of the directivity pattern in order to enhance the desired source of audio. By steering, we mean re-positioning of the main lobe to a specific angle. In other words, beamforming techniques try to direct its *look direction* to the source of interest.

3. Data Simulation

As we study the influence of distortion introduced by the room acoustics, the need for both the clean recordings and their noisy versions is obvious. However, there is no available multichannel SRE data. We, therefore, decided to use the dataset released for NIST Year 2010 Speaker Recognition evaluations [12]. From 9 evaluation conditions, condition 1 has been chosen in which all the trials are from the same microphone in training and test. The choice made due to the clarity of audio because it will undergo further processing. In this case, the recordings were captured by the close-talking microphones.

Condition 1 evaluation results form our best-case baseline. The far-field data were acquired by the simulation of room acoustics. For the simulation, we used the *Room impulse response Generator* tool [13] by E. Habets. It is based on *image method* [14] which outputs room impulse responses for a given pair of source and receiver in arbitrarily large room with specified characteristics of walls. The convolution of the obtained impulse responses with the original signals from the condition 1 set leads to set of parallel audio data as if it was recorded in that room (with simplifications introduced by simulation method). To increase diversity of conditions, two differently sized rooms were simulated – 4 × 4.5 × 3 m and 8 × 10 × 5 m. We will refer to them as "room" and "hall", respectively. The setup of the smaller room is shown in Figure 3.

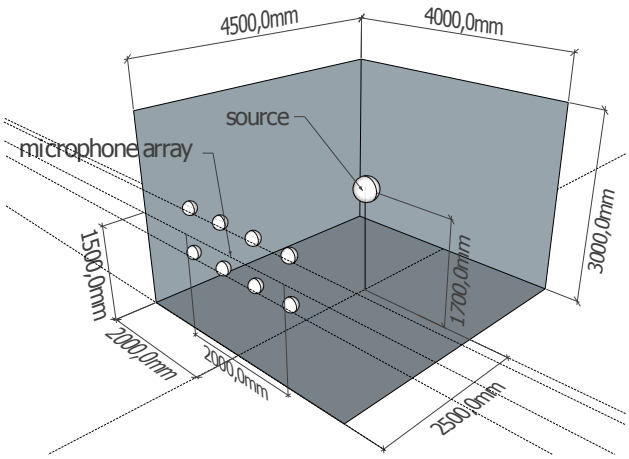

Figure 3. Model of the simulated room (4 × 4.5 × 3 m).

For the adaptation of the SRE system, we needed data that match new conditions. The SRE training datasets are typically huge (> 1000 hours). Thus, a real recording would not be feasible in a reasonable time period. The training data were acquired by simulation of rooms with random dimensions and the placement of the microphone array was changing as well.

4. Preprocessing Multichannel Data and Speaker Recognition System Description

4.1 Delay-and-sum

Delay-and-sum is the simplest and the most intuitive beamforming method. It makes use of the fact that the propagation delay causes the original sound wave to arrive at different instants of time to each microphone. When the *time difference of arrival* (TDOA) is known, the signals recorded by microphones can be shifted accordingly. Like this the desired signal is aligned in

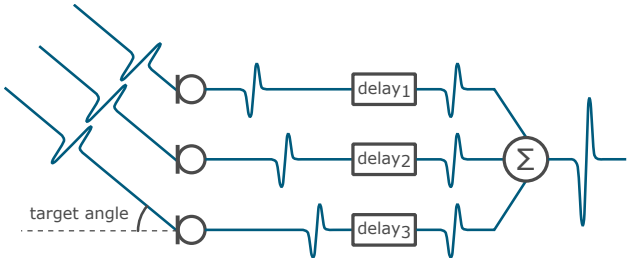

Figure 4. Principle of beamforming. Signals that come from the direction of interest are aligned and hence amplified.

time, while the other components included in the audio remain unaligned, and get attenuated. Figure 4 shows the principle.

We wrote the script that performs delay-and-sum beamforming in MATLAB. For the computation, the knowledge of TDOAs is needed. Since the spatial locations of all microphones and the source are known, delays can be calculated beforehand. However, in a real-world scenario, speaker position is unknown and may differ in time. To tackle this problem, multiple approaches have been developed [15]. Many of them are based on the alignment of signals according to their similarity. A straightforward option is cross-correlation, but it deals badly with reverberation. We used more universal variant – *generalized cross-correlation* (GCC) with PHAT weighting – that was shown to be less prone to errors caused by the echo [15]. Hence, beamforming works in the frequency domain. The script divides input recordings into overlapping frames whose length is 500 ms. The delay between the reference microphone and another microphone is given as the number of samples that correspond to the maximum of GCC – the inverse Fourier transform of PHAT weighted product of one spectrum and the second conjugate spectrum.

4.2 Speaker recognition system

MFC coefficients of dimension 60 are extracted from recordings in 10 ms steps. Such features were used for training of the UBM, which comprises 2048 components. I-vector dimensionality is 600 and they are further projected to 200-dimensional space using *linear discriminant analysis* (LDA). The latent variables in PLDA are of the same dimensionality.

5. Experiments

The first experiment is aimed at quantifying the deterioration when the original SRE system is used for far-field recordings. The next part is dedicated to methods applied in order to diminish this deterioration. In all the experiments, the two rooms described in section 3 are used: "room" and "hall".

Table 1. Accuracy of the SRE system for the clean and noisy test data in terms of equal error rate [%] (the lower the better).

	Clean	Room	Hall
Female	2.07	10.82	10.62
Male	0.61	6.93	6.47

To get the baseline results, we evaluated NIST 2010 SRE condition 1 (clean data) using the original system. The accuracy of the same system was also evaluated for the test data that simulate room the conditions. The outcome of this phase in terms of equal error rate (EER) is shown in Table 1. We can see significant degradation. In the next stage, we will explore various possibilities to make this difference smaller.

In the first approach, the whole system remained the same and a microphone array was used instead of single microphone. In our case, delay-and-sum was applied to a planar microphone array consisting of 8 microphones (Figure 3). The effect is shown in Figure 5 (blue bars). In all the subsequent experiments, we also need to pay attention to the accuracy of the system on the clean test data. In this first case it did not change, as the delay-and-sum is a preprocessing applicable to multichannel data only, hence it has no effect on the clean single-channel data.

The following experiments focus on adaptation of the SRE system. At first, we re-trained the i-vector extractor (see Figure 1 for its placement in the processing pipeline). The training dataset was augmented by one copy of the original training data, in which we added distortions that simulate random rooms (section 3). Moreover, this synthetic data were processed by delay-and-sum. The results are displayed in Figure 5 (yellow bars). It would be appropriate to perform yet another experiment in which no beamforming would be applied to the training data. Due to time constraints, we decided to skip it and the results should be worse than with the beamforming.

Next, we experimented with the PLDA retraining (see Figure 1 for the role of PLDA), while the original i-vector extractor was used and no beamforming was performed. As in the case of i-vector extractor retraining, the training dataset was extended with the randomly simulated copy. Note that the original training data for PLDA and i-vector extractor differ. The outcome of this experiment is shown in Figure 5 (green bars).

Based on presented results, it seems beneficial to perform PLDA retraining so that it can learn a new vari-

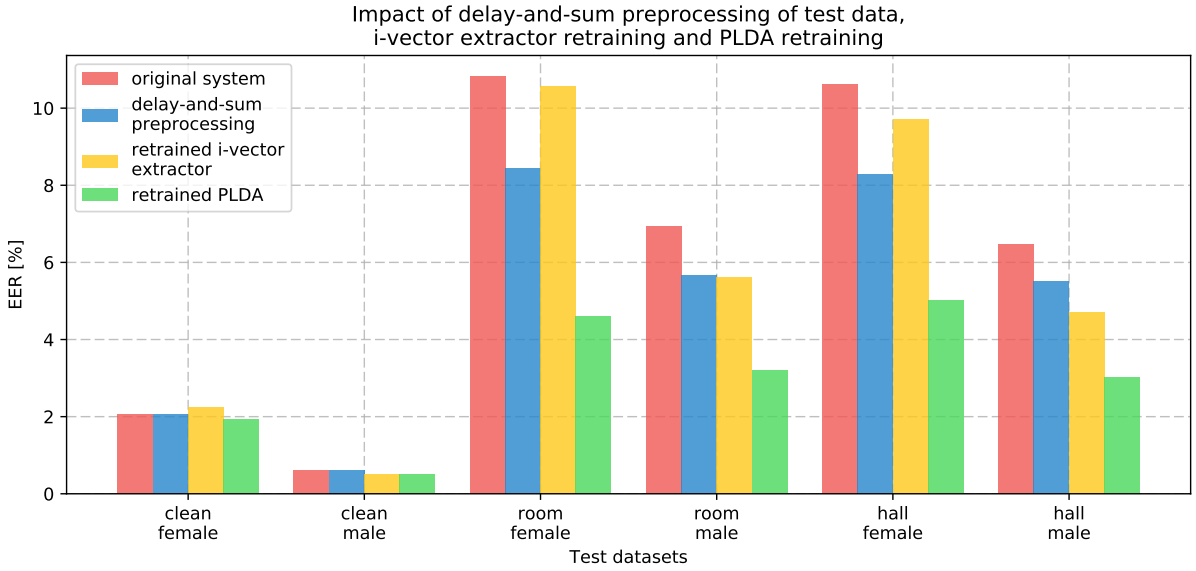

Figure 5. Impact of separately applied delay-and-sum preprocessing, i-vector extractor retraining and PLDA retraining in terms of EER for different test conditions.

ability introduced by the room conditions even though it is partially suppressed by preceding delay-and-sum. The extended training dataset also enhanced the robustness as the EERs on the clean test data decreased. Delay-and-sum itself evidently helps to improve the accuracy, as well.

Until now, one single change at a time was considered – only beamforming or only i-vector retraining or only PLDA retraining. The combination of the individual techniques was the next step. To compare their performance, we introduce a measure called *recovery from error* (RFE)

$$RFE(x,g,r) = \left(1 - \frac{x - a_{clean}(g)}{a_{room}(g,r) - a_{clean}(g)}\right) \cdot 100, \quad (2)$$

where g and r specify test conditions – g is gender, r is a type of room. $a_{clean}(g)$ is the accuracy of the original system for the clean test data of gender g in terms of EER. $a_{room}(g,r)$ is analogous, but test data from room r are considered (single far-field microphone). The symbol x refers to the new accuracy obtained by applying modifications of the system or preprocessing, whereas conditions match those specified by g and r. When RFE equals zero, it means that new result does not help to recover from errors. On the other hand, 100 % is achieved in case that a particular technique helped the system to reach the original accuracy. Table 2 summarizes impacts of combinations of recovery approaches presented beforehand. We can see that in almost every case the combination of techniques outperforms any of the separate techniques. The shortcuts in the table have the following meanings:

DS delay-and-sum preprocessing,
ivec_r i-vector extractor retraining,
PLDA_r PLDA retraining.

Table 2. Performance of enhancing techniques in terms of RFE [%] (the higher the better). In each column, the best recovery is highlighted.

	Room		**Hall**	
	Female	**Male**	**Female**	**Male**
DS	27.4	20.1	21.2	16.4
ivec_r	2.8	20.9	10.6	30.2
DS + ivec_r	27.4	34.4	33.2	32.8
PLDA_r	71.0	59.2	65.5	58.6
DS + PLDA_r	75.9	**68.0**	75.4	60.3
ivec_r + PLDA_r	71.1	54.5	70.0	61.2
DS + ivec_r + PLDA_r	**77.9**	63.2	**78.0**	**62.1**

6. Conclusions

In this paper, we have dealt with a topic of far-field speaker recognition. Different techniques that cope with the errors that are introduced by reverberation were discussed. Two approaches were outlined – preprocessing of multichannel data (from a microphone array) known as beamforming (delay-and-sum in our case) and adaptation of the system to simulated far-field data.

We have shown that delay-and-sum, i-vector extractor retraining and PLDA retraining are beneficial

when dealing with the far-field scenario. Based on the experiments, we concluded that the combination of the aforementioned methods leads to even greater accuracy improvements: up to 75 % recovery of the performance gap between close-talk microphone data and single far-field microphone data.

Even though a substantial improvement was achieved, there is still room for more improvement. In this work, only the delay-and-sum beamforming method was used, while more elaborate techniques exist, for example minimum variance distortionless response (MVDR) or generalized sidelobe canceller (GSC) [16]. Also, we worked with the simulated data, but the simulation neglects some of the physical principles of sound wave propagation. As a part of the future work, the correlation with real world data should be verified.

Acknowledgements

I would like to express my sincere thanks to my supervisor Honza Černocký for his valuable advices, shared expertise and positive attitude. I would also like to thank members of BUT Speech@FIT group, especially Oldřich Plchot, Ondřej Glembek and Pavel Matějka, for being willing to advise.

References

[1] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19(issue 4):788–798, 2011. ISSN: 15587916.

[2] Fang Zheng, Guoliang Zhang, and Zhanjiang Song. Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, vol. 16(issue 6):582–589, 2001. ISSN: 10009000.

[3] Ondřej Glembek. *Optimization of Gaussian Mixture Subspace Models and Related Scoring Algorithms in Speaker Verification*. PhD thesis, Brno University of Technology, Faculty of Information Technology, 2012.

[4] Jan Silovský. *Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvčího*. PhD thesis, Technická univerzita v Liberci, Fakulta mechatroniky, informatiky a mezioborových studií, 2011.

[5] Yu Zhang. Useful Derivations for i-Vector Based Approach to Data Clustering in Speech Recognition. 2011. http://people.csail.mit.edu/yzhang87/tech/Ivector.pdf.

[6] Simon J.D. Prince and James H. Elder. Probabilistic Linear Discriminant Analysis for Inferences About Identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. ISBN: 9781424416301.

[7] Patrick Kenny. Bayesian speaker verification with heavy-tailed priors. In *Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010.

[8] Niko Brümmer. EM for Probabilistic LDA. 2010.

[9] Aleksandr Sizov, Kong Aik Lee, and Tomi Kinnunen. Unifying Probabilistic Linear Discriminant Analysis Variants in Biometric Authentication. page 464.

[10] Kenichi Kumatani, John McDonough, and Bhiksha Raj. Microphone Array Processing for Distant Speech Recognition. *IEEE Signal Processing Magazine*, vol. 29(issue 6):127–140, 2012. ISSN: 10535888.

[11] Iain McCowan. Microphone Arrays : A Tutorial. 2001.

[12] The NIST Year 2010 Speaker Recognition Evaluation Plan, 2010. https://www.nist.gov/sites/default/files/documents/itl/iad/mig/NIST_SRE10_evalplan-r6.pdf.

[13] Emanuël A.P. Habets. Room Impulse Response Generator, September 2010. https://github.com/ehabets/RIR-Generator/blob/master/rir_generator.pdf.

[14] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, 1979. ISSN: 0001-4966.

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

[16] Mehrez Souden, Jacob Benesty, and Sofiène Affes. On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18(issue 2):260–276, 2010. ISSN: 15587916.