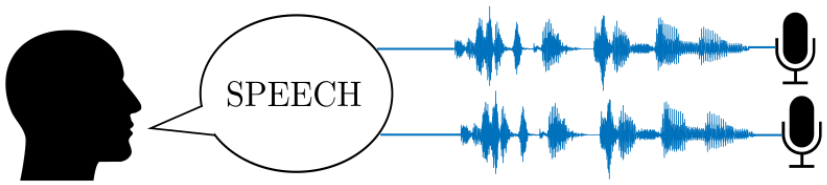
BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2016
http://excel.fit.vutbr.cz

Far-field speech recognition

Kateřina Žmolíková*



Abstract
The paper deals with the problem of speech recognition using distant microphones. The usage of distant microphones is often more convenient in real applications than using the close talking ones. However, this introduces the problem of noise and reverberation which degrades the accuracy of the speech recognition system. This problem can be reduced by using microphone arrays rather than single microphone. This paper explores the methods of processing of multichannel recordings to enhance the speech, thereby improving the speech recognition performance. To process the array and achieve noise reduction, two different methods (Delay-and-sum and Minimum variance distortionless response beamforming) are explored. For dereverberation, Weighted prediction error method is used. The methods are tested on three different noisy datasets (AMI, CHiME3 and REVERB). The results achieved on these tasks are comparable with the published ones. In this paper, we present the applied methods, analyze the results and discuss the modifications necessary for achieving the results.

Keywords: Speech recognition — Microphone arrays — Beamforming — Dereverberation

Supplementary Material: *N/A*

*xzmoli02@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Automatic speech recognition is a research field aiming at automatically translating spoken language into text. Through last fifty years it has evolved from recognizing small vocabulary of carefully pronounced words to recognizing spontaneous speech from many languages. Nowadays, the accuracy of the state-of-the-art systems is already sufficient for being used in real-world applications such as intelligent personal assistants or in-car systems.

However, the translation of speech technologies into real-world applications gives rise to many problems which were not present in small artificial tasks. In real tasks it is often more convenient to use far-field rather than close talking microphones. This introduces lots of distortions causing significant performance degradation. In such far-field setting, possible solution is to use microphone array rather than a single microphone to reduce the problem. The usage of multiple microphones enables to use spatial information during the pre-processing stage which can significantly help to separate the speech signal from the surrounding noise.

The objective of this work is to explore existing methods for processing of microphone arrays in the context of far-field speech recognition, implement chosen techniques and evaluate them on the speech recognition task. This paper will first summarize the problems arising in far-field speech recognition (Section 2), then introduce used methods to deal with these problems (beamforming methods in Section 3 and dereverberation methods in Section 4), present three noise-

robust tasks to evaluate them (Section 5) and finally discuss the achieved results (Section 6).

2. Far-field speech recognition

To define the problem and motivate the need for the presented techniques, the main concepts of speech recognition are introduced and the influence of distortions in the far-field scenario is discussed.

2.1 Automatic speech recognition

The task of speech recognition system is to transcribe speech signal into the sequence of words which were spoken. The main obstacle which makes this task difficult is huge variability of speech — signals corresponding to the same words can differ largely due to numerous factors such as speaking style or context variations. A successful speech recognition system must be able to deal with this variability.

The basic architecture of the system which has proven to be effective in solving this problem is depicted in Figure 1. The input signal is typically first processed by *feature extraction* component that transforms it into a representation more suitable for the rest of the system. The resulting features are then exploited by *acoustic model* which incorporates the knowledge about acoustics and phonetics. The useful information about language which is being recognized is represented by the *language model*. The parameters of both of these models are mostly trained on a large corpora of annotated speech and text. The *hypothesis search* component combines the outcomes of these models to provide the final result of the recognition.

This system is designed to be invariant to the variability present in the speech signal. Thus it should perform satisfactory even in the far-field scenario. However, it is often not the case. The following text will discuss the problems emerging when distant microphones are used to capture the speech signal.

2.2 Far-field scenario

In case speech signal is captured by a distant microphone, it contains high amount of ambient distortions. This causes unwanted variability leading to a mismatch between the trained acoustic model and the tested speech. As a consequence, the performance of the speech recognition system significantly drops.

The distortions which cause such degradation in the far-field scenario can be classified into two main categories

additive noise Distortions caused by additional sound sources in the environment. The severity of the additive noise can be measured by Signal-to-noise ratio (SNR), which is the ratio of the speech and noise power, usually expressed in decibels.

reverberation The effect of repeated reflections of the original signal on the walls and objects in the room. It can be characterized by reverberation time T_{60} which refers to the time needed for the reverberation to decay by 60 dB. More thorough description of reverberation will be provided in section 4.

In [1], Seltzer shows the effect of additive noise and reverberation on the error rate of the speech recognition system. For SNRs less than 20 dB the word error rate increases sharply. When the SNR level reaches about 10dB (speech power being 10 times higher than noise power) the error is already at 50%, making the system almost unusable. The same applies to experiments with reverberation where increasing the reverberation time up to 0.5 seconds (which is in the range of a typical office or home environment) again makes the word error rate to reach 50%. This motivates the need for methods dealing with noise and reverberation. The common solution is the use of microphone arrays.

3. Beamforming

The most popular class of methods to process signals from microphone arrays is beamforming. These methods aim to combine the signals received at individual microphones in such a way that a signal coming from particular direction is enhanced while signals coming from the other directions are attenuated. This section describes two widely used methods — *Delay-and-sum* (DS) and *Minimum variance distortionless response* (MVDR) beamforming.

3.1 Delay and sum

Delay-and-sum (DS) is a simple and straight-forward method to perform beamforming. It uses the fact that microphones situated at different spatial positions receive the same source signal with different delays. Moreover, the delays vary with the direction from which the source signal is coming. If we know the delays corresponding to the desired direction, we can use them to shift the signals. The shift operation aligns the desired signal in all channels, while the signals coming from different directions remain unaligned. We can then simply average all such shifted signals which will cause attenuation of the signals from unwanted directions. The entire process is summarized in Figure 2.

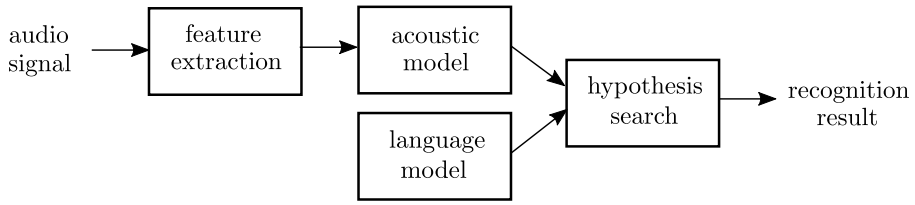


Figure 1. Scheme of speech recognition system.

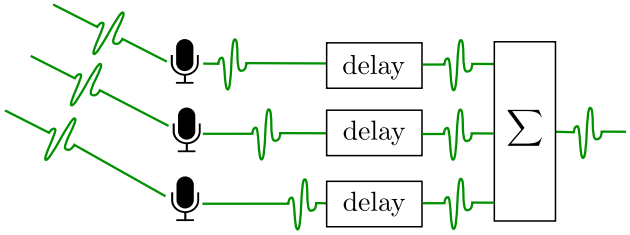


Figure 2. Scheme of Delay and sum beamformer.

To use this method, we need to know the delays of the source signal as received by each microphone. This can be either known apriori (from the microphone array architecture and the position of the source) or, in more common case, we need to estimate the delays from the received signals. The most popular techniques for time delay estimation are based on cross correlation. Cross correlation reflects the level of similarity of two signals as a function of their relative delay. With some reasonable assumptions, we can expect that the delay leading to the maximum of this function matches the true delay at which the speech signal arrived at the microphones. For more detailed overview of used time delay estimation methods, see [2].

3.2 Minimum variance distortionless response

Despite its simplicity, the Delay-and-sum is a very common choice for beamforming and often performs sufficiently well. However, the main drawback of the DS method is, that it estimates its parameters considering only the position of the desired source and not the positions of the interfering sounds. This gives rise to Minimum variance distortionless response beamformer (MVDR) which explicitly aims to minimize the effect of noise.

To achieve the maximum noise reduction, it uses an estimate of noise covariance matrix which represents how correlated the noise signals are between the microphones. The knowledge of these correlations provides the information about the directions of the noise sources and enables the beamformer to supress the signals coming from these directions. More detailed description of the MVDR beamformer is beyond the scope of this paper and the reader is referred to [3, 4].

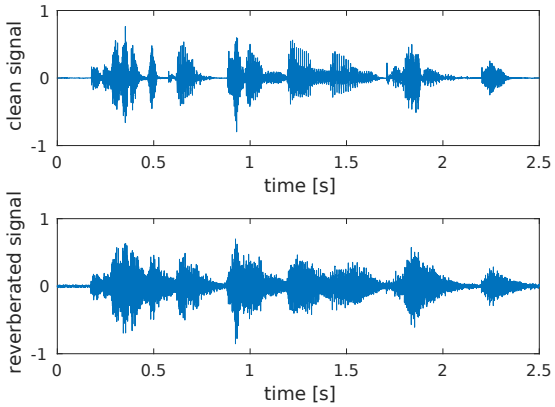


Figure 3. Comparison of clean and reverberated signal.

4. Dereverberation

Speech signal captured by distant microphones contains not only the additive noise, but also reflections of the same signal from walls and other objects — effect known as reverberation. This type of distortion has very distinctive properties causing many conventional noise reduction methods to fail. In [5], Habets points out that traditional beamforming algorithms become ineffective in the presence of reverberation. As a consequence, it is necessary to use special methods to dereverberate the input signal.

This section summarizes the properties of reverberation that make the problem of reverberant speech so challenging. Next, it introduces an effective dereverberation method — *Weighted prediction error*, that was used in experiments presented in this paper.

4.1 Properties of reverberation

Figure 3 shows the comparison of clean and reverberant speech in time domain. It can be seen that reverberation causes smearing in time which has the effect of masking between subsequent phonemes. Perceptually, these effects lead to speech sounding “distant” and “echoic”.

The characteristic features of reverberation that complicate the process of recognizing reverberant speech are the following [6]:

high non-stationarity As reverberation is a result of

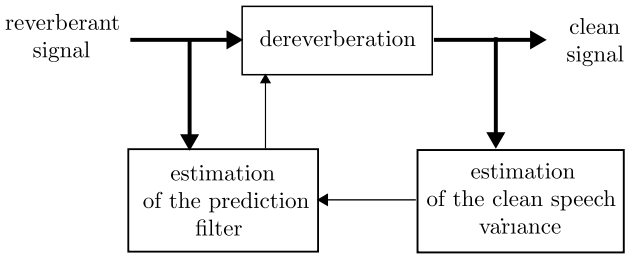


Figure 4. Weighted prediction error method.

filtering delayed speech signal, its characteristics vary rapidly. This makes most noise reduction techniques unsuitable for this task because they typically assume stationary noise.

long-term effects The models of speech typically work on the basis of short (about 20 millisecond) frames. The effect of reverberation spans across multiple frames creating long-term relationships between the feature vectors. Since the widely used models in speech recognition system assume conditional independence between feature vectors, they are ineffective for the modeling of reverberation.

4.2 Weighted prediction error

There are multiple classes of approaches to solve these problems. In this work, we focused on Weighted prediction error method (WPE) because it has been shown successful in recent evaluations [7, 8], it can effectively make use of multichannel signals and it can be easily coupled with beamforming techniques.

WPE [9, 10] uses the fact that the reverberation component is dependent on the previous samples of the speech signal. It aims to predict the reverberant component using the previous observations and substract this prediction from the reverberant signal. This leads to an estimate of clean speech signal.

This procedure is based on the classical linear prediction. An important difference from the classical linear prediction (as used for example in speech coding) is that WPE assumes a model of the target speech signal with time-varying variance. This variance needs to be estimated together with the prediction filter which leads to an iterative algorithm alternating between the estimation of the filter coefficients and the estimation of the properties of the clean speech signal. The procedure is summarized in Figure 4.

5. Data

To evaluate and analyze the implemented methods, we decided to use three datasets — AMI, CHiME3 and REVERB, that cover noisy and reverberant conditions and contain the recordings obtained using microphone arrays.

5.1 AMI

The AMI Meeting Corpus [11] is a collection of meetings recorded in three standardized meeting rooms. The recordings are obtained using 12 microphones — a headset microphone per participant and an 8-element circular microphone array. All meetings are in English, though mostly spoken by non-native speakers. The total amount of recorded data is about 100 hours, which is partitioned into train, dev and eval sets following [12]. This makes about 78 hours of speech for training, 9 hours for development and 9 hours for evaluation.

5.2 CHiME3

The CHiME3 dataset was recorded for the 3rd CHiME Challenge for Speech Separation and Recognition [13]. It includes recordings of people speaking in real environments including cafe, street junction, public transport and pedestrian area. Recordings are obtained using tablets with six-channel microphone array. In addition to the real recordings, part of the data was obtained by artificially mixing clean speech data with noisy backgrounds.

The training set comprises 1600 utterances from four speakers speaking in real environments and 7138 simulated utterances from 83 speakers. The development set contains 1640 real and 1640 simulated utterances from four speakers (which do not overlap with speakers from training data). The evaluation set contains 1320 real and 1320 simulated utterances from four other speakers.

5.3 REVERB

The REVERB dataset was recorded for REVERB challenge [14] aimed at evaluation of speech enhancement and recognition in reverberant environments. The recordings were obtained with a distant microphone array in reverberant rooms with a limited amount of stationary noise. For all data, one-microphone, two-microphone and eight-microphone versions were available — here, we focused on eight microphone recordings.

The training data were obtained by mixing clean data from WSJCAM0 [15] dataset with room impulse responses and noise signals measured in real rooms. The development set consists of data from four rooms. For three of them, the data were simulated, for the fourth room the recordings were real. The evaluation set consists of the same environments as the dev set, but with different speakers and different positions in the rooms.

6. Experiments

In this section, we will present results obtained by applying the above mentioned methods on the task of noise robust speech recognition. First, we will summarize the achieved results on all three datasets. Next, we will discuss the setup and the modifications of the beamforming methods that were performed to obtain the best results.

For building the speech recognition systems in this work, we used Kaldi speech recognition toolkit [16]. The tested methods were implemented in Matlab. The accuracy of the speech recognition systems is measured using Word error rate (WER), which refers to the percentage of words that were recognized incorrectly.

6.1 Overall results

Table 1 shows the overall results on all three datasets achieved with Delay-and-sum and MVDR methods for beamforming and WPE method for dereverberation. The results are compared with the baseline error rates achieved by using the signal from a single channel only (the channel giving the best results).

The results clearly show the advantage of using microphone arrays instead of relying on a single microphone in the far-field scenario. Moreover, the results also show that the MVDR outperfoms the basic Delay-and-sum beamformer on both CHiME3 and AMI datasets. For the REVERB dataset, we did not use MVDR beamforming as this dataset contains small amount of additive noise, so we did not expect the results to notably improve.

We can also see that WPE is helping to achieve better results on all datasets. Figure 5 shows the comparison of distorted and dereverberated utterance using the WPE method. It can be seen that WPE "shortens" the effect of the reverberation as expected.

Table 1. Overall results in terms of word error rates on three datasets.

	CHiME3		AMI	REVERB	
	simu	real		simu	real
1-best channel	19.85	25.19	63.3	12.01	30.07
DS	15.99	18.15	58.35	8.06	23.30
MVDR	15.69	17.67	57.93	-	-
DS + WPE	15.46	17.81	58.39	**5.65**	**15.48**
MVDR + WPE	**15.37**	**17.32**	**57.89**	-	-

6.2 Beamforming implementation details

The theoretical foundations of both Delay-and-sum and MVDR beamformers lead to quite simple formulas for the processing of input signals. However, in real applications, practical issues often arise requiring additional modifications. Here we describe the most notable improvements applied on top of the basic algorithms to achieve the aforementioned results. Table 2 then shows the individual benefits due to each of these modifications on CHiME3.

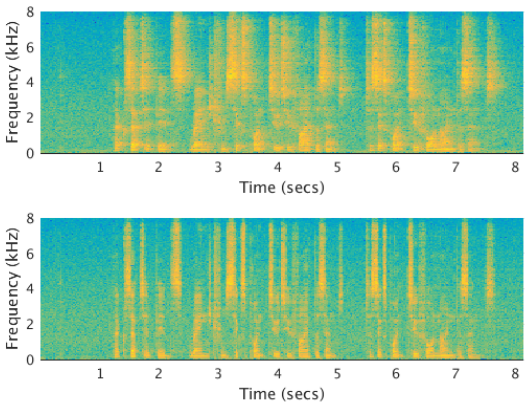


Figure 5. Comparison of spectrogram of reverberant (upper) and dereverberated (bottom) speech using WPE method.

weighting channels Since signals in certain channels might be more corrupted than others, it is beneficial to weigh the signals during the sum operation accordingly. Our computation of the weights is based on cross-correlation measure between the channels following [17].

skipping unreliable frames Some parts of the input signals may not be reliable for estimation of time delays. This applies mainly for very noisy parts or segments not containing any speech. We detected the silence parts using robust neural-network based Voice activity detector [18, Section 4] and further filtered the segments where the cross-correlation was too low.

fractional delays For small microphone arrays, the true delay between the microphones may be smaller than the sampling period of the signals. In such case, the estimation of the delays can be improved by delaying the signals by fractions of one sample. This can be done by interpolating the signal between the samples.

PHAT correlation weighting The time delay estimation method can be made more robust by incorporating frequency-domain weighting of the cross-correlation measure. This technique was suggested in [19].

Table 2. The modifications of Delay-and-sum and MVDR beamforming methods and their improvements of CHiME3 dataset.

	CHiME3	
	simu	real
no modifications	19.81	23.15
weighting channels	17.94	21.64
skipping unreliable delays	18.63	21.35
fractional delays	18.10	22.50
PHAT weighting	18.13	22.27
all together	**15.99**	**18.15**

7. Conclusions

In this paper we have presented the methods for microphone array processing for the far-field speech recognition task. Particularly, we focused on two beamforming techniques (Delay-and-sum and Minimum variance distortionless response). These techniques enable to combine signals from multiple microphones to reduce the noise. Additionally, we used Weighted error prediction method to deal with the problem of reverberation.

We evaluated the methods on three noise-robust tasks (AMI, CHiME3 and REVERB) and achieved significant improvement over the single microphone case which corresponds to the published state-of-the-art. We showed that the MVDR method outperforms simple Delay-and-sum and that the results can be further improved using dereverberation techniques. We discussed the modifications of these methods leading to better results.

One possible way to further improve the accuracy of the presented methods is to train the entire front-end including beamforming together with the acoustic model of the speech recognition system. Recently, several paper were published in this area [20, 21] showing promising results. We will explore this approach in the future.

Acknowledgements

I would like to thank my supervisor Honza Cernocky for many valuable suggestions and feedback.

References

[1] M.L. Seltzer. *Microphone Array Processing for Robust Speech Recognition*. PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 2003.

[2] J. Chen, J. Benesty, and Y. Huang. Time delay estimation in room acoustic environments: An overview. *EURASIP Journal on Applied Signal Processing*, pages 170–170, 2006.

[3] K. Kumatani, J. McDonough, and B. Raj. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Processing Magazine*, 29(6):127–140, 2012.

[4] 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*, 18(2):260–276, 2010.

[5] E. Habets. *Single- and multi-microphone speech dereverberation using spectral enhancement*. PhD thesis, Technische Universiteit Eindhoven, 2007.

[6] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, 29(6):114–126, 2012.

[7] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, et al. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016.

[8] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, Ch. Yu, W.J. Fabian, M. Espi, T. Higuchi, et al. The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 436–443, 2015.

[9] T. Yoshioka, T. Nakatani, M. Miyoshi, and H.G. Okuno. Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):69–84, 2011.

[10] 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*, 20(10):2707–2720, 2012.

[11] J. Carletta. Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41:181–190, 2007.

[12] P. Swietojanski, A. Ghoshal, and S. Renals. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 285–290, 2013.

[13] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green. The PASCAL CHiME speech separation and recognition challenge. pages 621–633, 2013.

[14] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas. The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–4, 2013.

[15] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 81–84, 1995.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The Kaldi speech recognition toolkit. In *IEEE workshop on automatic speech recognition and understanding*, 2011.

[17] X. Anguera. *Robust Speaker Diarization for Meetings*. PhD thesis, UPC Barcelona, 2006.

[18] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselý, and P. Matejka. Developing a speech activity detection system for the darpa rats program. In *INTERSPEECH*, pages 1969–1972. ISCA, 2012.

[19] Ch. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 320–327, 1976.

[20] T. Sainath, R. J. Weiss, K.W. Wilson, A.Narayanan, and M. Bacchiani. Factored spatial and spectral multichannel raw waveform cldnns. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[21] 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 *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.