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

Introduction

This work deals with far-field speaker recognition.
We demonstrate and investigate:

- the degree of degradation of the state-of-the-art i-vector based speaker recognition system on reverberant data,
- PLDA re-training,
- preprocessing techniques: dereverberation, beamforming,
- development of SR system of competitive accuracy in far-field settings.

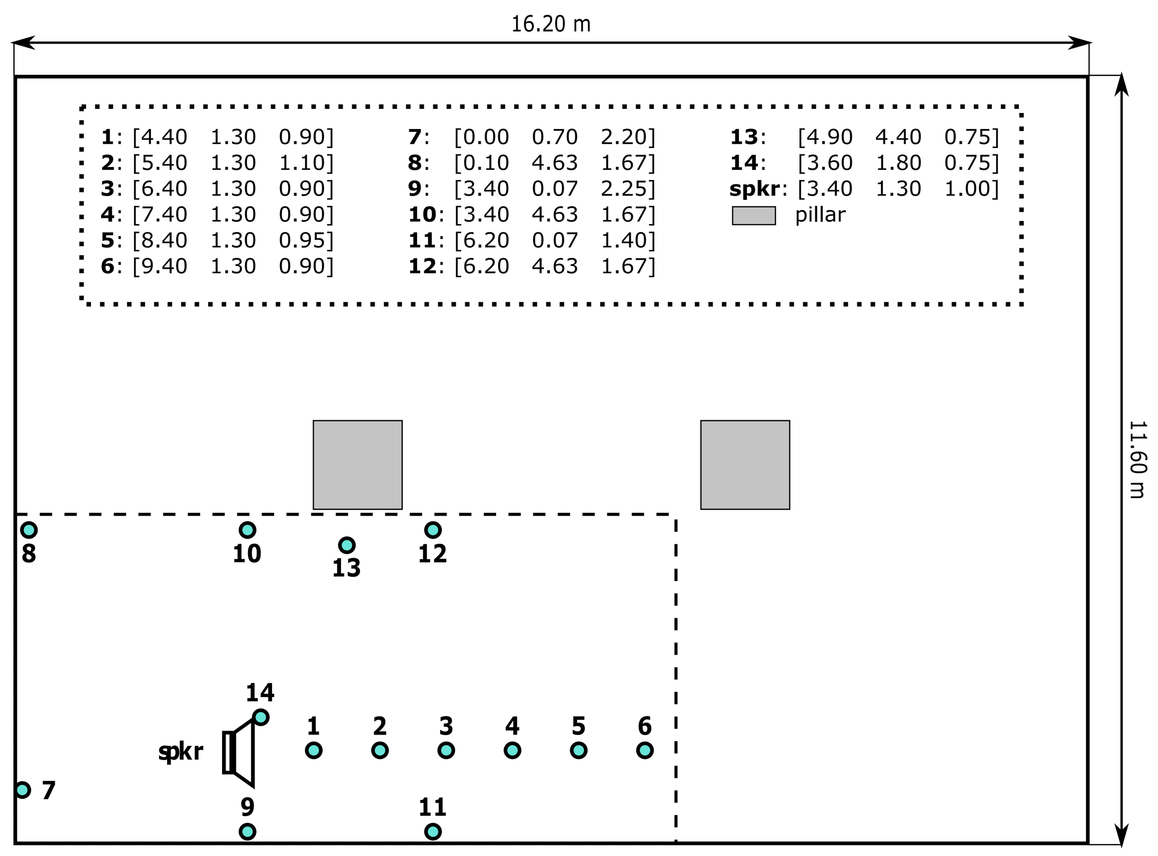
Experimental setup

Test dataset

For this work, a subset of data released for NIST Year 2010 Speaker Recognition evaluations (SRE) was retransmitted.

- duration of recordings: 3 min and 8 min

	Number of recordings	Number of speakers
Female	459	150
Male	473	150



Floor plan of the room in which the retransmission took place. Coordinates are in meters and lower left corner is the origin. 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.

Speaker recognition system

- Mel-frequency cepstral coefficients: 60-dimensional (including Δ and $\Delta\Delta$)
- Cepstral Mean and Variance Normalization: 3s window
- GMM-UBM: 2048 components
- i-vectors: 200-dimensional (projected by LDA from 600-dimensional space)
- Probabilistic Linear Discriminant Analysis

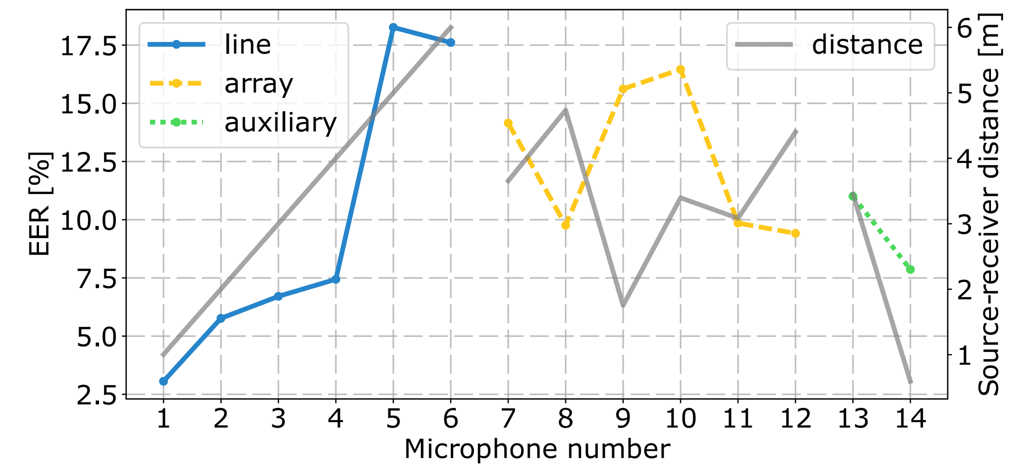
Experiments

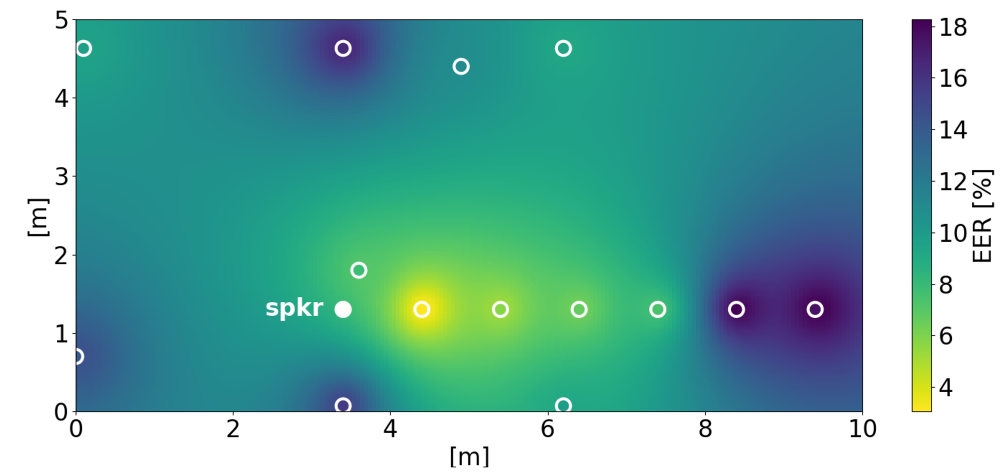
All the results of experiments are expressed in equal error rates (EER). For convenience, we show only female test data results. The baseline accuracy – **2.52% EER** – was obtained on clean test data before the retransmission.

Adverse effects of distance on speaker recognition

The test data captured by individual microphones were evaluated with the original system. *line*: inter-microphone distance of 1 m (microphones 1...6); *array*: large microphone array (microphones 7...12); *auxiliary*: remaining sensors (microphones 13, 14).

- distance-accuracy correlation does not hold for the *array*
- the result of a directivity pattern and local acoustic conditions





System adaptation

adapt_simu system

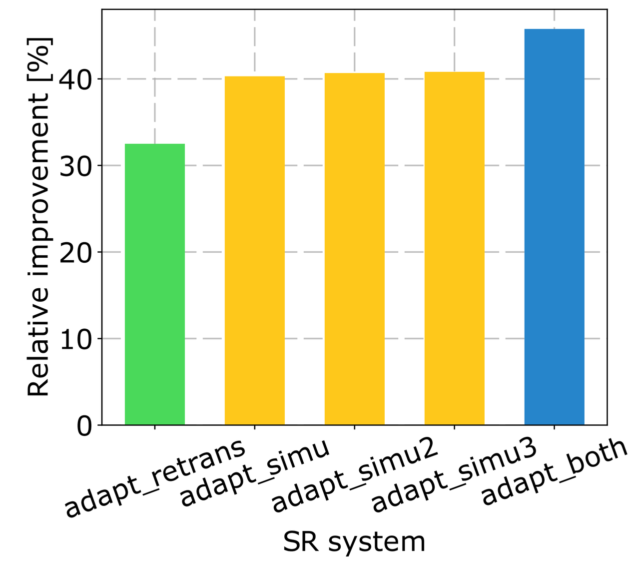
- 1 copy of the original training data (86680)
- 1 copy of the simulated data (86680)
  - image method simulation [1]
  - random dimensions of rooms, positions of microphones

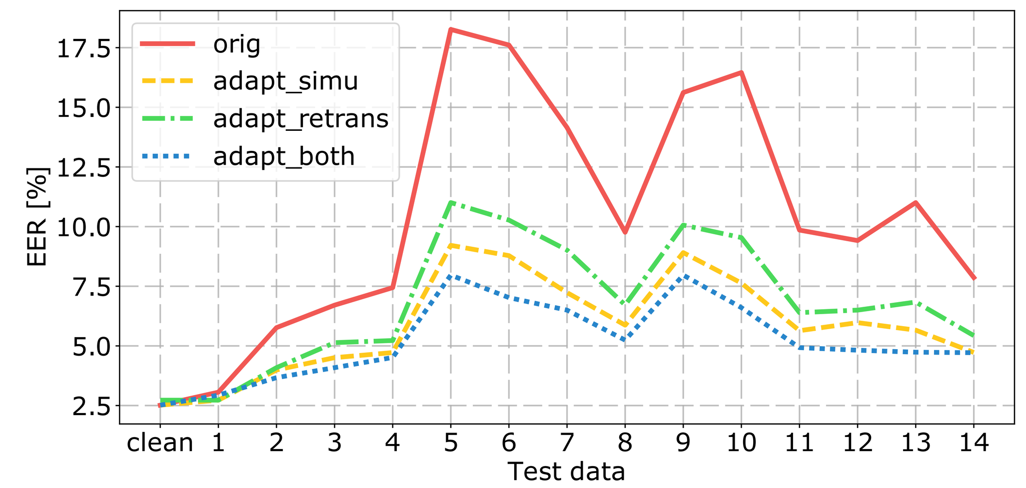
adapt_retrans system

- 1 copy of the original training data (86680)
- part of retransmitted data (6524)
  - jackknifing

adapt_both system

- concatenated condition





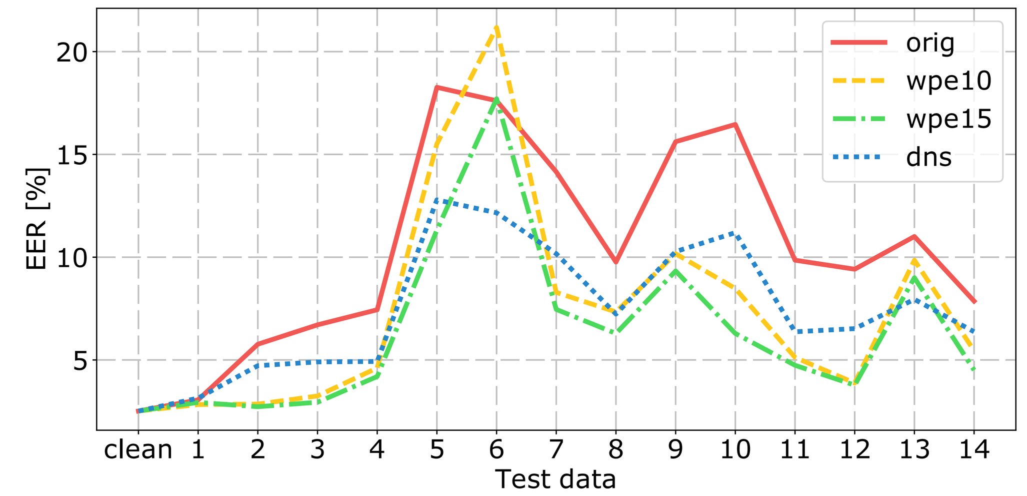
Dereverberation

WPE

- weighted prediction error [2]
- wpe10: 10 filter coefficients, wpe15: 15 filter coefficients

DNS

- denoising/dereverberation autoencoder
- input: a central frame of a log-magnitude spectrum with a context of +/− 15 frames
- output: enhanced central frame



References

[1] 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.
[2] 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.
[3] 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.
[4] 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.

Microphone array experiments

Beamforming

- microphones 7...12

DS: delay-and-sum

- GCC-PHAT for TDOA estimation

MVDR: minimum variance distortionless response

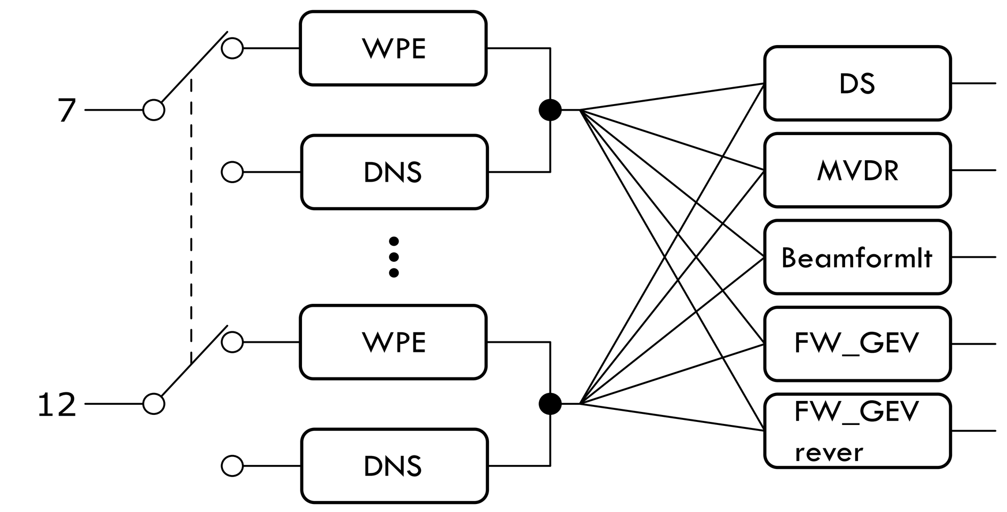
- diffuse noise field assumption

BeamformIt: weighted delay-and-sum + additional processing [3]
FW_GEV: generalized eigenvalue beamformer [4]

- feed-forward NN for PSD masks estimation
- *FW_GEV_rever*: simulated reverberant training data

Test data		Original system	Simulated data adapt.
Clean		2.52	2.52
Reverberant	Best	9.42	5.64
	Worst	16.46	8.91
DS		14.15	9.01
MVDR		13.62	7.44
BeamformIt		9.43	6.08
FW_GEV		10.07	5.56
FW_GEV_rever		**7.54**	**4.93**

Combinations of techniques



Test data		Original system	Simulated data adapt.	Dereverb. data adapt.
DNS	Best	6.37	5.03	4.09
	Worst	11.19	8.28	7.45
WPE	Best	3.88	3.67	3.56
	Worst	10.17	9.22	7.87
DNS + DS		9.33	6.71	6.18
DNS + MVDR		9.45	6.50	5.75
DNS + BeamformIt		8.49	6.84	6.19
DNS + FW_GEV		7.36	5.66	5.24
DNS + FW_GEV_rever		**6.29**	**4.30**	**4.50**
WPE + DS		6.18	6.08	5.66
WPE + MVDR		6.18	5.03	4.93
WPE + BeamformIt		5.03	4.30	4.09
WPE + FW_GEV		2.83	**2.73**	**2.62**
WPE + FW_GEV_rever		**2.73**	2.83	2.73

Simulated data adapt.: equals to adapt_simu system
Dereverb data adapt.: the same as adapt_simu, an additional appropriate dereverberation technique was applied to the simulated portion of the data

The best results

2.52% baseline accuracy, clean test data

9.42% best-performing microphone from the microphone array

2.62% the best result we achieved