# Deep Beamforming and Data Augmentation for Robust Speech Recognition: Results of the 4th CHiME Challenge

*Tobias Schrank, Lukas Pfeifenberger, Matthias Zöhrer, Johannes Stahl, Pejman Mowlaee, Franz Pernkopf*

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
lukas.pfeifenberger@alumni.tugraz.at,
{tobias.schrank,matthias.zoehrer,johannes.stahl,pejman.mowlaee,pernkopf}@tugraz.at

## Abstract

Robust automatic speech recognition in adverse environments is a challenging task. We address the 4th CHiME challenge [1] multi-channel tracks by proposing a deep eigenvector beamformer as front-end. To train the acoustic models, we propose to supplement the beamformed data by the noisy audio streams of the individual microphones provided in the real set. Furthermore, we perform data augmentation by modulating the amplitude and time-scale of the audio. Our proposed system achieves a word error rate of 4.22% on the real development and 8.98% on the real evaluation data for 6-channels and 6.45% and 13.69% for 2-channels, respectively.

## 1. Background

This report describes our proposed ASR system for the 6- and 2-channel task of the 4th CHiME challenge. The proposed modifications of the baseline system are:

- As multi-channel front-end we employ an optimal multi-channel Wiener filter, which consists of an eigenvector GSC beamformer and a single-channel postfilter. Both components depend on a speech presence probability mask, which we learn using a deep neural network (DNN).
- In addition to the beamformed signals we use noisy multi-channel real data to train the acoustic model of the ASR, i.e. we perform *multi-channel* training.
- We perform data augmentation by modulating the signal amplitude (volume perturbation) and time-scale modifications (speed perturbation).
- We perform sequential language model rescoring using (gated) RNNs.
- We combine multiple systems with a lattice-based approach which uses minimum Bayes risk decoding.

A detailed introduction of the individual components and relevant literature are provided in the next section.

## 2. Robust Multi-Channel ASR System

Figure 1 shows the block diagram of the proposed multi-channel ASR system including the data augmentation and multi-channel training of the recognizer. Each processing step is detailed in the following sections.

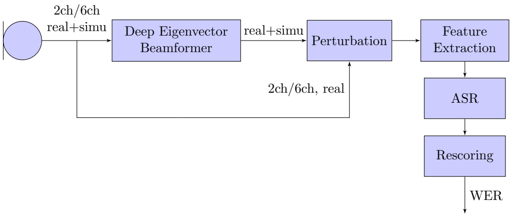

Figure 1: System overview.

### 2.1. Deep Eigenvector Beamformer

As multi-channel speech enhancement front-end we employ a *deep eigenvector beamformer*, which consists of a generalized sidelobe canceller (GSC) beamformer [2–6], followed by a single-channel postfilter. The GSC consists of a steering vector $\boldsymbol{F}$, a blocking matrix $\boldsymbol{B}$, and an adaptive interference canceller, such that: $\boldsymbol{W} = \boldsymbol{F} - \boldsymbol{B}\boldsymbol{H}_{AIC}$. The GSC block diagram is given in Figure 2. The steering vector $\boldsymbol{F}$ has to model the *acoustic transfer functions* (ATFs) from the speaker to the microphones [7]. Usually this is done by a *direction of arrival* (DOA) estimation. However, this method does not include the complex propagation paths present in the CHiME4 data. Therefore we use the dominant eigenvector of the speech PSD matrix $\hat{\boldsymbol{\Phi}}_{SS}$ as steering vector $\boldsymbol{F}$, such that the beamformer is directed towards the speech source in signal subspace. This allows the beamformer to account for early echoes and reverberation of the speaker signal [7–9]. Hence, we refer to this beamformer as *eigenvector GSC* (EV-GSC).

Using the steering vector $\boldsymbol{F}$, the blocking matrix is given as $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}^H$. The adaptive interference canceller $\boldsymbol{H}_{AIC}$ is learned using an adaptive NLMS filter [10]. The single-channel postfilter consists of a real-valued gain mask $G = \frac{\xi}{1+\xi}$, which is obtained from the SNR $\xi$ at the beamformer output. It is given as $\xi = \frac{\boldsymbol{W}^H \hat{\boldsymbol{\Phi}}_{SS} \boldsymbol{W}}{\boldsymbol{W}^H \hat{\boldsymbol{\Phi}}_{NN} \boldsymbol{W}}$. The SNR depends on both the speech and noise PSD matrices, which are estimated using a time and frequency dependent *speech presence probability* $p_{SPP}$.

We use a DNN to learn $p_{SPP}$ from the dominant eigenvector of the PSD matrix of the noisy inputs. As we are operating in the frequency domain, each frequency bin $k$ is assigned to a kernel as shown in Figure 3. The feature vector $\boldsymbol{x}_k$ for each kernel consists of the cosine distance between the eigenvectors of 5 consecutive frames. This introduces some context-

This work was supported by the LEAD project, the Austrian Science Fund (FWF) under the project number P25244-N15 and P27803-N15 and the K-Project ASD. Furthermore, we acknowledge NVIDIA for providing GPU computing resources.

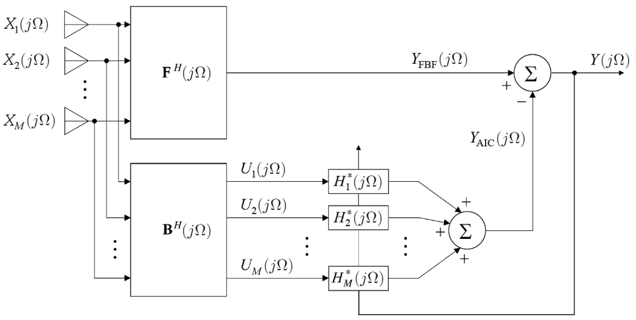

Figure 2: GSC beamformer

sensitivity into our model. The DNN of each kernel uses a hybrid model with a generative and a discriminative component [11]. The generative component consists of two autoencoder layers, which perform unsupervised clustering of the input data $\boldsymbol{x}_k$. The autoencoder kernels operate independently for each frequency bin. We used 20 neurons in the first layer, and 10 neurons in the second layer. The discriminative component consists of a regression layer which fuses the activations of all autoencoder kernels, in order to exploit information which is distributed across the frequency. The regression layer predicts the $K$ output labels $p_{SPP}(\boldsymbol{x}_k)$). Figure 3 illustrates the kernelized DNN used in our system.

For more details on the EV-GSC beamformer and the kernelized DNN, we refer the reader to [12]. We use the same architecture for the 2ch and 6ch track, as the training data is the same for both tracks.

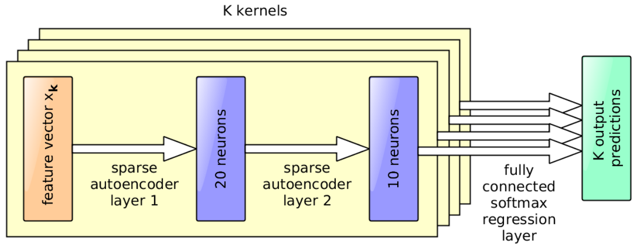
K kernels
feature vector $x_k$
sparse autoencoder layer 1
20 neurons
sparse autoencoder layer 2
10 neurons
fully connected softmax regression layer
K output predictions

Figure 3: Kernelized DNN to estimate the speech presence probability $p_{SPP}$

### 2.2. ASR

The ASR system employs a hybrid DNN architecture which is implemented with the Kaldi toolkit [13]. We do not only use the beamformed data for training but add the noisy channels of the real data (except for channel 2 which faces backwards). With this *multi-channel training (MC)* we can both compensate for the small amount of training data and make the acoustic model less sensitive to noise that might be left over in the evaluation data. In the evaluation stage we still use only the beamformed signals.

The GMM system uses 13 MFCCs and their deltas and delta-deltas. The DNN uses 40 fMLLR features extracted from this GMM system. For the DNN the data is augmented with speed-perturbed copies of the original data. Additionally, the data is volume-perturbed for greater robustness (*pert*). The DNN is then generatively pre-trained using restricted Boltzmann machines. The DNN has 6 hidden layers and is trained with a state-level minimum Bayes risk (*sMBR*) criterion. The results which have been obtained in this way are then rescored with a Kneser-Ney smoothed 5-gram model (*5-gram*), a recurrent neural network language model (*RNNLM*) and a gated RNNLM (*GRNNLM*). The two RNNLMs consist of a single hidden layer with 300 and 500 neural units, respectively.

We perform system combination by first combining the lattices of the system with perturbed training data (*pert*), the system with multi-channel training (*MC*) and the system that uses both (*MC + pert*). We then decode the resulting lattices with an sMBR criterion.

## 3. Experimental Evaluation

Table 1 shows the results of our systems for the 6-channel and 2-channel tasks of the 4th CHiME challenge. For each data set the best score for a single system and for system combination is in boldface. Due to time constraints we report only those results for the 2-channel task which uses the system architecture that we have found to be optimal for the 6-channel task ($\mathrm{S_C}$). Therefore the following comparison focuses on the 6-channel task.

On average over the test sets, our proposed EV-GSC beamformer of S2 performs 2% WER better than the baseline *BeamformIt* beamformer of S1, i.e. 7.95% WER vs. 9.98% WER for the RNNLM-rescored DNN. However, this performance improvement is the least pronounced for the real evaluation data. Data augmentation through speed perturbation and volume perturbation (*pert*) of S3 results in an improvement of .74% WER on average, i.e. 7.20% WER vs. 7.95% WER. Multi-channel (MC) training of S4 leads to an improvement of 0.80% WER on average, i.e. 7.15% WER vs. 7.95% WER. Both multi-channel training and amplitude and time-scale perturbation (MC+pert) of S5 results in an improvement of 1.19% WER on average, i.e. 6.75% WER vs. 7.95% WER. Further rescoring with the gated RNNLM leads to a small improvement of 0.04% WER. The best results for 6-channels are achieved by a combination of systems S3, S4, and S5 as S6. In particular, we obtain a WER of 8.98% and 7.02% on the real and simulated test set, respectively.

Table 2 shows the individual results for each environment of our best system for the 6- and 2-channel track. For both systems, performance on the real evaluation data set is considerably worse for BUS than for any other environment.

## 4. References

[1] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” *Computer Speech and Language*, to appear.

[2] O. Hoshuyama, A. Sugiyama, and A. Hirano, “A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters,” *IEEE Transactions on Signal Processing*, vol. 47, no. 10, Oct. 1999.

[3] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin–Heidelberg–New York: Springer, 2008.

[4] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Transactions on Signal Processing*, vol. 49, no. 8, Aug. 2001.

[5] W. Herbordt and W. Kellermann, “Analysis of blocking matrices for generalized sidelobe cancellers for non-stationary broadband signals,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, May 2002.

Table 1: Average WER (%) for the tested systems.

| Track | System | | | | Dev | | Test | |
|---|---|---|---|---|---|---|---|---|
| | Tag | ASR | Data | BF | real | simu | real | simu |
| 2ch | $S_A$ | GMM | – | EV-GSC | 14.16 | 15.13 | 26.33 | 24.12 |
| | $S_B$ | GMM | MC | EV-GSC | 13.41 | 15.36 | 23.46 | 23.49 |
| | $S_C$ | DNN | MC + pert | EV-GSC | 9.38 | 11.33 | 17.92 | 18.10 |
| | | +sMBR | | | 9.24 | 10.91 | 17.16 | 17.46 |
| | | +5-gram | | | 7.63 | 9.60 | 15.29 | 15.81 |
| | | +RNNLM | | | 6.66 | 8.54 | 14.02 | 14.46 |
| | | +GRNNLM | | | **6.45** | **8.29** | **13.69** | **14.33** |
| 6ch | S1 | GMM | – | beamformit | 12.78 | 14.87 | 23.13 | 23.06 |
| | | DNN | | | 9.41 | 10.43 | 17.26 | 17.14 |
| | | +sMBR | | | 8.33 | 9.21 | 15.72 | 15.88 |
| | | +5-gram | | | 6.91 | 7.96 | 13.75 | 13.63 |
| | | +RNNLM | | | 5.99 | 7.16 | 12.21 | 12.42 |
| | | +GRNNLM | | | 6.03 | 7.21 | 12.07 | 12.50 |
| | S2 | GMM | – | EV-GSC | 11.21 | 11.92 | 23.41 | 16.13 |
| | | DNN | | | 8.32 | 8.32 | 17.36 | 11.75 |
| | | +sMBR | | | 7.37 | 7.52 | 15.55 | 10.83 |
| | | +5-gram | | | 6.01 | 6.14 | 14.05 | 9.35 |
| | | +RNNLM | | | 5.14 | 5.48 | 12.60 | 8.56 |
| | | +GRNNLM | | | 5.16 | 5.51 | 12.64 | 8.35 |
| | S3 | DNN | pert | EV-GSC | 7.82 | 7.96 | 16.13 | 11.01 |
| | | +sMBR | | | 6.83 | 6.86 | 14.34 | 10.16 |
| | | +5-gram | | | 5.66 | 5.76 | 12.78 | 8.70 |
| | | +RNNLM | | | 4.71 | 5.13 | 11.53 | 7.44 |
| | | +GRNNLM | | | 4.74 | **5.05** | 11.45 | **7.34** |
| | S4 | GMM | MC | EV-GSC | 11.05 | 11.77 | 19.65 | 15.93 |
| | | DNN | | | 8.15 | 7.94 | 14.38 | 11.37 |
| | | +sMBR | | | 7.30 | 7.49 | 13.38 | 10.56 |
| | | +5-gram | | | 5.82 | 6.17 | 11.55 | 9.51 |
| | | +RNNLM | | | 4.96 | 5.27 | 10.23 | 8.14 |
| | | +GRNNLM | | | 4.86 | 5.29 | 10.08 | 8.06 |
| | S5 | DNN | MC + pert | EV-GSC | 7.65 | 8.03 | 13.53 | 10.89 |
| | | +sMBR | | | 6.81 | 7.24 | 12.50 | 10.01 |
| | | +5-gram | | | 5.53 | 6.08 | 10.94 | 8.57 |
| | | +RNNLM | | | **4.65** | 5.35 | 9.63 | 7.38 |
| | | +GRNNLM | | | 4.66 | 5.28 | **9.54** | 7.38 |
| | S6 | combination | | EV-GSC | **4.22** | **4.73** | **8.98** | **7.02** |

Table 2: WER (%) per environment for the best system.

| Track | Envir. | Dev | | Test | |
|---|---|---|---|---|---|
| | | real | simu | real | simu |
| 2ch | BUS | 8.35 | 7.24 | 19.46 | 9.28 |
| | CAF | 5.78 | 10.80 | 13.41 | 16.92 |
| | PED | 4.23 | 5.86 | 12.07 | 15.00 |
| | STR | 7.45 | 9.25 | 9.81 | 16.12 |
| 6ch | BUS | 5.25 | 3.79 | 13.72 | 4.20 |
| | CAF | 3.98 | 5.99 | 7.12 | 7.73 |
| | PED | 2.79 | 3.58 | 7.31 | 8.29 |
| | STR | 4.85 | 5.56 | 7.79 | 7.86 |

[6] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, Nov. 2004.

[7] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin–Heidelberg–New York: Springer, 2008.

[8] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin–Heidelberg–New York: Springer, 2006.

[9] 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, Jul. 2007.

[10] P. Vary and R. Martin, *Digital Speech Transmission*. West Sussex: Wiley, 2006.

[11] M. Zöhrer, R. Peharz, and F. Pernkopf, "Representation learning for single-channel source separation and bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2398–2409, 2015.

[12] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Dnn-based speech mask estimation for eigenvector beamforming," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017, submitted.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *ASRU*, 2011.