

# Speaker Adapted Beamforming for Multi-Channel Automatic Speech Recognition

Tobias Menne, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition, Computer Science Department,  
RWTH Aachen University, Aachen, Germany  
{menne, schlueter, ney}@cs.rwth-aachen.de

## Abstract

This paper presents, in the context of multi-channel ASR, a method to adapt a mask based, statistically optimal beamforming approach to a speaker of interest. The beamforming vector of the statistically optimal beamformer is computed by utilizing speech and noise masks, which are estimated by a neural network. The proposed adaptation approach is based on the integration of the beamformer, which includes the mask estimation network, and the acoustic model of the ASR system. This allows for the propagation of the training error, from the acoustic modeling cost function, all the way through the beamforming operation and through the mask estimation network. By using the results of a first pass recognition and by keeping all other parameters fixed, the mask estimation network can therefore be fine tuned by retraining. Utterances of a speaker of interest can thus be used in a two pass approach, to optimize the beamforming for the speech characteristics of that specific speaker. It is shown that this approach improves the ASR performance of a state-of-the-art multi-channel ASR system on the CHiME-4 data. Furthermore the effect of the adaptation on the estimated speech masks is discussed.

**Index Terms:** robust ASR, multi-channel ASR, speaker adaptation, acoustic beamforming, CHiME-4

## 1. Introduction

The performance of automatic speech recognition (ASR) systems has shown significant improvements over the last decade. Those have especially been driven by the utilization of deep learning techniques [1]. Nevertheless the performance of systems dealing with realistic noisy and far-field scenarios is still significantly worse than the performance of close talking systems on clean recordings [2, 3]. Multi-channel ASR systems are often used in those scenarios to improve recognition robustness. In these systems the effect of noise, reverberation and speech overlap is mitigated by utilizing spatial information through beamforming [4].

Usually beamforming is done in a separate preprocessing step before applying the ASR system to the enhanced signal, which is obtained from the output of the preprocessing [5]. A general formulation for beamforming is the filter-and-sum approach [6, 7], where the single channels are summed up after applying a separate linear filter to each one. Usually those filters are derived such that an objective criterion on the signal level, such as signal-to-noise ratio (SNR), is optimized. Popular approaches are the delay and sum (DAS) [4], minimum variance distortionless response (MVDR) [8] and generalized eigenvalue (GEV) [9] beamforming methods. Most systems submitted to the CHiME and REVERB challenges [10, 11, 12] follow one or more of these approaches.

The objective used to optimize the preprocessing thus differs from the objective of the acoustic model training. Even before the introduction of deep neural network (DNN) hybrid systems in ASR, the optimization of the preprocessing towards the goal of speech recognition was proposed e.g. in [13]. The success of deep learning also motivated the integration of the beamforming operation into the acoustic model. E.g. in [14, 15] the filters of the filter-and-sum beamforming are estimated by a neural network based on input features derived from the multi-channel input signal. Even learning the complete multi-channel preprocessing, starting from the raw time signal, has been shown to work [16, 17, 18]. The advantage of those approaches is, that the preprocessing is not optimized for a proxy measure like SNR at the output of the beamformer, but directly towards the criterion for acoustic model training. But thus far, a very large amount of training data is necessary to obtain satisfying performance with those approaches.

Lately the performance of statistically optimal beamformers was improved by using neural networks to estimate speech and noise masks, which are then used to compute the beamforming vectors [19, 20, 8]. This approach has worked well for many submissions to the 4<sup>th</sup> CHiME challenge [5, 21, 22]. One problem of that approach is the need for target masks in the mask estimator training, which usually requires stereo data (the noisy and its respective clean signal) to create the target masks for training. Since this type of data is much more difficult to collect than only the noisy data, training of the mask estimator is usually done on simulated signals, which can lead to a mismatch between training and test data. To solve this problem, the authors of [23] proposed to integrate the mask based, statistically optimal beamforming with the acoustic modeling of the ASR system. This enables the propagation of the training error all the way through the acoustic model and the mask estimator network in the preprocessing. Therefore the mask estimator can be trained based on the training criterion of the acoustic model training.

In this paper, the approach of integrating the mask based, statistically optimal beamformer with the acoustic model is utilized to adapt the mask estimation to the speech characteristics of a speaker of interest in a two pass recognition approach.

The rest of the paper is organized as follows. An overview of the integrated system is given in Section 2. Furthermore an alternative approach to [23] for the propagation of the gradients through the eigenvalue problem of the beamformer is presented. Section 3 describes the experimental setup of a state-of-the-art system for the CHiME-4 speech recognition task followed by the experimental results in Section 4.

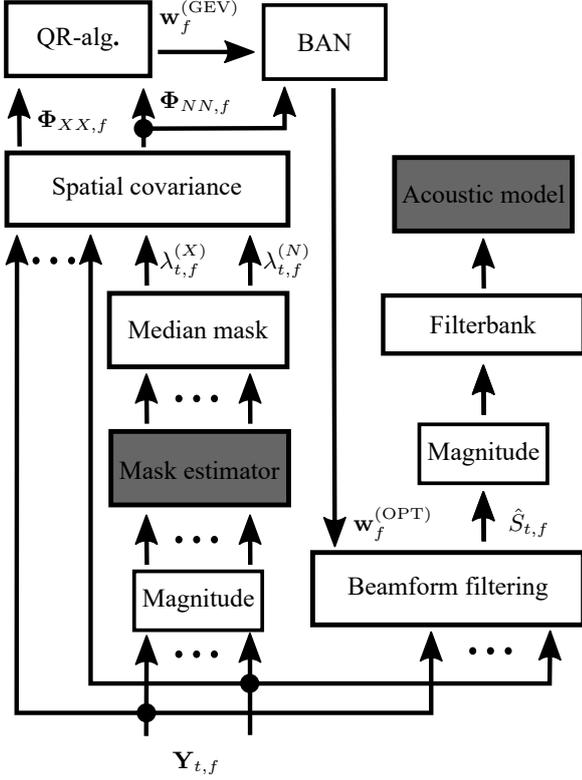


Figure 1: Overview of the integrated system. The grey blocks indicate modules with trainable parameters.

## 2. System overview

The system used in this work integrates the acoustic beamformer, usually called front-end, with the acoustic model of the ASR system, usually called back-end, very similarly to the integration described in [23]. Figure 1 gives an overview of the integrated system.  $\mathbf{Y}_{t,f}$  is the input in the short-time Fourier transform (STFT) domain, recorded from an array of  $M$  microphones. It consists of a speech component  $\mathbf{X}_{t,f}$  and a noise component  $\mathbf{N}_{t,f}$ :

$$\mathbf{Y}_{t,f} = \mathbf{X}_{t,f} + \mathbf{N}_{t,f} \quad (1)$$

Where  $\mathbf{Y}_{t,f}, \mathbf{X}_{t,f}, \mathbf{N}_{t,f} \in \mathbb{C}^M$ ,  $t$  is the time frame index and  $f$  is the frequency bin index.

The main difference to the system introduced in [23] will be described in Section 2.3, whereas the acoustic beamformer and acoustic model are described in Sections 2.1 and 2.2, respectively. During a first pass decoding a hidden Markov model (HMM)-state sequence  $s_1^T$  is obtained for the input signal  $\mathbf{Y}_{1,1}^{T,F}$ , where  $T$  and  $F$  are the number of time frames and frequency bins of the signal. Section 2.4 describes the utilization of the state sequence to adapt the acoustic beamformer to a certain speaker.

### 2.1. GEV beamformer

The main purpose of the front-end is to denoise the input signal. Here this is achieved by acoustic beamforming [6, 7]:

$$\hat{\mathbf{S}}_{t,f} = \mathbf{w}_f^H \cdot \mathbf{Y}_{t,f} \quad (2)$$

Where  $\hat{\mathbf{S}}_{t,f} \in \mathbb{C}$  is an estimate of the speech component, obtained by applying the beamforming vector  $\mathbf{w}_f \in \mathbb{C}^M$ .  $(\cdot)^H$  denotes the Hermitian transpose.

For this work we use the GEV beamformer with blind analytic normalization (BAN), as described in [9] and which is also used in [23]. The beamforming vector of the GEV beamformer is derived by maximizing the *a posteriori* SNR:

$$\mathbf{w}_f^{(\text{GEV})} = \underset{\mathbf{w}_f}{\operatorname{argmax}} \frac{\mathbf{w}_f^H \Phi_{XX,f} \mathbf{w}_f}{\mathbf{w}_f^H \Phi_{NN,f} \mathbf{w}_f} \quad (3)$$

Where  $\Phi_{XX,f}$  and  $\Phi_{NN,f}$  are the spatial covariance matrices of speech and noise, respectively. This results in the generalized eigenvalue problem

$$\Phi_{XX,f} \mathbf{W} = \Phi_{NN,f} \mathbf{W} \Lambda \quad (4)$$

with  $\mathbf{w}_f^{(\text{GEV})}$  being the eigenvector corresponding to the largest eigenvalue.

The spatial covariance matrices  $\Phi_{\nu\nu,f}$  for  $\nu \in \{X, N\}$  are computed by applying a mask  $\lambda_{t,f}^{(\nu)}$  to the recorded multi-channel signal  $\mathbf{Y}_{t,f}$

$$\Phi_{\nu\nu,f} = \frac{1}{\sum_{t=1}^T \lambda_{t,f}^{(\nu)}} \sum_{t=1}^T \lambda_{t,f}^{(\nu)} \mathbf{Y}_{t,f} \mathbf{Y}_{t,f}^H \quad (5)$$

A mask estimating neural network is used to estimate  $\lambda_{t,f}^{(X)}$  and  $\lambda_{t,f}^{(N)}$ . For both, speech and noise, one mask is estimated for every channel,  $\lambda_{t,f}^{(\nu)}$  is then computed as the median mask, which contains the element-wise median of the channel dependent masks, as described e.g. in [19].

The BAN post-filter, as described in [9], is a frequency dependent scaling of the GEV beamforming vector, such that the final beamforming vector used here is:

$$\mathbf{w}_f^{(\text{OPT})} = w_f^{(\text{BAN})} \cdot \mathbf{w}_f^{(\text{GEV})} \quad (6)$$

With  $w_f^{(\text{BAN})} \in \mathbb{C}$  being the scaling factor described in [9].

### 2.2. Acoustic model

The acoustic model is a bidirectional long short-term memory (BLSTM) hybrid model using log-mel filterbank features as input. Apart from the features, the training pipeline is the same as for the speaker independent model described in [5].

### 2.3. Beamformer integration into acoustic model

Training of the integrated system presented in Figure 1 is done according to standard error back propagation. The gradient computation for the propagation through the acoustic model, feature extraction, linear filtering of the beamformer, BAN and mask estimator network are straight forward. To propagate the gradient through the computation of the principal eigenvector of

$$\Phi_f = \Phi_{NN,f}^{-1} \Phi_{XX,f} \quad (7)$$

as required for computing the beamforming vector  $\mathbf{w}_f^{(\text{GEV})}$  according to Equation 4, the derivatives of the eigenvalue problem w.r.t.  $\Phi_{NN,f}$  and  $\Phi_{XX,f}$  are derived in [24] and used in [23].

In contrast, here the principal eigenvector of Equation 7 is approximated by applying the QR-algorithm as presented in [25]. A matrix  $A_k$  is decomposed by the QR-decomposition

into a product of a unitary matrix  $Q_k$  and an upper triangular matrix  $R_k$ :

$$A_k = Q_k R_k \quad (8)$$

With  $k$  being the iteration index,  $A_{k+1}$  is then computed as

$$A_{k+1} = R_k Q_k \quad (9)$$

It is shown in [25], that  $A_K$  converges to an upper triangular matrix as  $K \rightarrow \infty$ . The diagonal of  $A_K$  then contains the eigenvalues of  $A_0$  and  $\prod_{k=0}^K Q_k$  contains the respective eigenvectors. This QR-algorithm is used here to approximate the principal eigenvector of  $\Phi_f$  by setting

$$A_0 = \Phi_f \quad (10)$$

The algorithmic differentiation of the QR decomposition is outlined in [26] and applied here in the error back propagation.

#### 2.4. Speaker adaptation of mask estimator

After a first pass recognition an optimal sequence of HMM states  $s_e^T$  is obtained from the decoding process for each of the evaluation segments of the speaker of interest. Those alignments are then used as training targets for an adaptation training of the integrated system. Of the system shown in Figure 1, only the parameters of the mask estimator are adjusted in the adaptation training. The parameters of the remaining pipeline are kept fixed, such that only the mask estimator network is tuned towards optimizing the cost function of the integrated system. Therefore the mask estimator and thus the computation of the beamforming vector are optimized for the speech characteristics of the speaker of interest.

Even though this work is using the GEV beamformer with BAN, it is noteworthy that the proposed speaker adaptation method is equally applicable to the mask based MVDR beamformer that is presented in [20], by changing the initialization of  $A_0$  in Equation 10 and omitting the BAN.

### 3. Experimental setup

The proposed speaker adaptation scheme for the acoustic beamformer is evaluated on the data of the CHiME-4 speech recognition task [11]. The CHiME-4 dataset features real and simulated 16 kHz, multi-channel audio data recorded with a six channel microphone array arranged around a tablet device. Based on the 5k WSJ0-Corpus recordings and simulations have been done with four different kinds of real-world background noise. The training set contains approximately 18 h of data per channel recorded from 87 different speakers. Results are provided for the real development and real evaluation set of the 6-Channel track. Both sets contain audio of 4 speakers each, of which 2 are male and 2 are female, with no overlap between development and evaluation set. The amount of data per speaker is approximately 0.7 h in the development set and around 0.5 h in the evaluation set.

The acoustic model used in the experiments is a BLSTM network, with 5 layers and 600 units per layer. Different to the system in [5], the input features are 80 dimensional log-mel filterbank features computed in the STFT domain employing a blackman window with a window size of 25 ms and a frame shift of 10 ms. The input features are unnormalized, but a linear layer with 80 units, employing batch normalization, was added as a first layer to the network. This results in a marginally better baseline system over the one described in [5]. The initial training of the acoustic model is done as described in [5], where

at first alignments for the training set are computed on the data of the close talking microphone by using a Gaussian mixture model (GMM)-HMM trained only on the data of the close talking microphones of the training set. Those alignments can then be used for all other channels, since the data is recorded sample synchronized. The training of the BLSTM acoustic model is done by using the unprocessed audio data of the single channels. This has been demonstrated to be beneficial in many submissions to the 3<sup>rd</sup> and 4<sup>th</sup> CHiME challenge e.g. in [27].

The mask estimator network used in the experiments is similar to the one described in [19]. It consists of a BLSTM layer with 256 units followed by two fully connected layers with 512 units and ReLU activations and another fully connected layer with 402 units and sigmoid activation. Thus the resolution of the estimated masks in frequency is lower than described in [19]. This is due to the adjustment of the dimensions of the masks to the discrete Fourier transform (DFT) size of the feature extraction pipeline of the ASR system used here. The input of the mask estimation network is the magnitude spectrum of a single channel. The output of the network is the concatenation of the noise mask and the speech mask. During decoding the outputs of the different channels, of one utterance, are grouped and the median masks are calculated. Those are then applied to all channels to estimate the spatial covariance matrices as described in Section 2.1. The initial mask estimation network is trained on the simulated training data as described in [19]. In contrast to [19], only the provided baseline configuration of the simulation is used and no additional data augmentation is done. The number of iterations of the QR-algorithm described in Section 2.3 is fixed to 5.

The decoding is done with the 5-gram language model provided as a baseline language model with the CHiME-4 dataset. In a post processing step a recurrent neural network (RNN) language model lattice rescoring is done. The RNN language model is a 3 layer long short-term memory (LSTM) with high-way connections. Details about the language model and lattice rescoring can be found in [5].

In addition to the acoustic beamforming described in Section 2.1, the baseline beamforming algorithm of the CHiME-4 task (BFIT) is used to provide baseline results. Apart from the beamforming algorithm, the exact same pipeline as described above is used.

The hyper-parameters for the speaker adaptation training such as the learning rate were tuned on the development set and applied to the evaluation set.

## 4. Experimental results

#### 4.1. Baseline systems

Table 1 shows an overview of the experimental results. It shows, that using the GEV front-end described in Section 2.1 yields an improvement of about 20% - 30% relative over the baseline system with the BFIT front-end. Joint training of the GEV front-end and acoustic model further improves the performance another 5% relative. Those results are in line with the results reported in [23]. When comparing the mask output of the mask estimator before and after joint training only minor differences in the masks can be observed. This is in line with the suggestion of the authors of [23], that a majority of the performance increase stems from the adaptation of the acoustic model towards the specific front-end.

Table 1: Average WER (%) for the described systems for different stages of the integrated training.

System				Dev	Eval
System id	Front-end	Joint training	Speaker adapted		
0	BFIT	-	-	4.36	7.17
1	GEV	+	+	3.46	5.18
2				3.32	4.84
3				<b>3.09</b>	<b>4.58</b>

#### 4.2. Speaker adapted beamforming

Table 1 shows an overall improvement of WER after speaker adaptation and Table 2 shows that improved performance is obtained for the majority of the speakers with an improvement in WER of up to 11 % and 15 % relative for single speakers of the evaluation and development set, respectively. Figure 2 shows an example of the estimated speech mask before and after speaker adaptation. It can be seen, that the speech mask after speaker adaptation shows a stronger emphasis on the fundamental frequency and the harmonics. This can be seen repeatedly between the time marks of 2 s and 3 s. At time mark 4 s a pattern of fundamental frequency and harmonics can be seen in the mask after adaptation, which is not present in the mask before adaptation and which can also hardly be spotted in the input signal or the clean signal. This could indicate an increased bias of the mask estimator towards this kind of pattern.

Table 2: WER (%) of separate speakers for the jointly trained system and the speaker adapted system

Sys. id	Dev				Eval			
	F01	F04	M03	M04	F05	F06	M05	M06
2	4.19	3.23	2.77	<b>3.07</b>	6.88	<b>4.09</b>	3.83	4.58
3	<b>3.55</b>	<b>3.20</b>	<b>2.48</b>	3.14	<b>6.35</b>	<b>4.09</b>	<b>3.38</b>	<b>4.48</b>

## 5. Conclusion

This work describes a method for speaker adaptation of mask based beamforming in a multi-channel ASR system. The basis of the adaptation method is the integration of the statistically optimal beamforming with the acoustic model to allow the back propagation of the training errors through the complete system, which has been previously introduced in [23]. Here an alternative solution for the back propagation of the errors through the computation of the beamforming vector, based on the QR-algorithm, is presented. The system is then used in a two pass approach to adapt the mask estimator to a speaker of interest during the decoding phase. It was shown that this adaptation method results in speech masks which show a stronger emphasis on the fundamental frequency and harmonics of the speaker. Furthermore a relative ASR improvement, for single speakers of the real evaluation data of the CHiME-4 ASR task, of up to 11 % relative was shown.

## 6. Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program grant agreement No. 694537. This work has also been supported by Deutsche Forschungs-

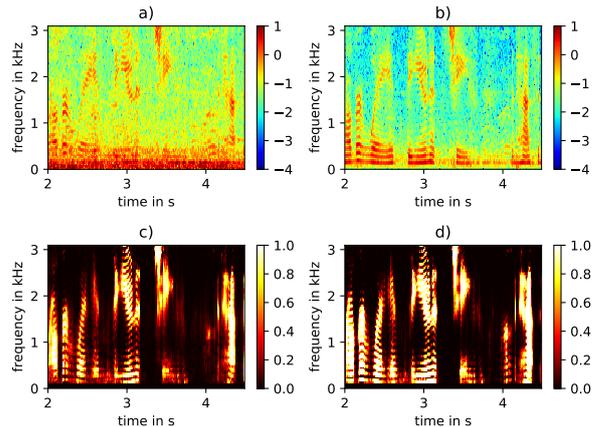


Figure 2: Two seconds snippet of the signal “F01\_421C0210\_BUS” of the development set starting at second 2 and showing the frequency range up to 3 kHz. a) log magnitude spectrum of the noisy signal recorded at channel 5 b) log magnitude spectrum of the signal recorded at the close talking microphone c) estimated speech mask of system 2 (jointly trained but before speaker adaptation) d) estimated speech mask of system 3 (after speaker adaptation)

gemeinschaft (DFG) under contract No. Schl2043/1-1 and European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 644283. The work reflects only the authors’ views and the European Research Council Executive Agency is not responsible for any use that may be made of the information it contains. The GPU cluster used for the experiments was partially funded by Deutsche Forschungsgemeinschaft (DFG) Grant INST 222/1168-1.

## 7. References

- [1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, J. Navdeep, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.
- [2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, Apr 2014.
- [3] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, H. Takaaki, and T. Nakatani, “Strategies for distant speech recognition in reverberant environments,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 4, pp. 60–74, Jul 2015.
- [4] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.
- [5] T. Menne, J. Heymann, A. Alexandridis, K. Irie, A. Zeyer, M. Kitzka, P. Golik, I. Kulikov, L. Drude, R. Schlüter, H. Ney, R. Haeb-Umbach, and A. Mouchtaris, “The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation,” in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016)*, San Francisco, CA, Sep. 2016, pp. 39–44.
- [6] E. Warsitz and R. Haeb-Umbach, “Acoustic filter-and-sum beamforming by adaptive principal component analysis,” in *Proc. IEEE*

- International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, PA, Mar 2005, pp. 797–800.
- [7] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr 1988.
  - [8] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, “Improved MVDR beamforming using single-channel mask prediction networks,” in *Proc. Interspeech*, San Francisco, CA, Sep 2016, pp. 1981–1985.
  - [9] E. Warsitz and R. Haeb-Umach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” *IEEE Transactions on audio, speech, and language processing*, vol. 15, pp. 1529–1539, Jun. 2007.
  - [10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, Dec 2015, pp. 504–511.
  - [11] 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 & Language*, vol. 46, pp. 535–557, Nov 2017.
  - [12] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, “The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2013, pp. 1–4.
  - [13] M. L. Seltzer, B. Raj, and R. M. Stern, “Likelihood-maximizing beamforming for robust hands-free speech recognition,” *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 489–498, Sep 2004.
  - [14] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar 2016, pp. 5745–5749.
  - [15] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, “Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar 2017, pp. 271–275.
  - [16] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, “Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, Dec 2015, pp. 30–36.
  - [17] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, “Factored spatial and spectral multichannel raw waveform cldnns,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar 2016, pp. 5075–5079.
  - [18] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, “Neural network adaptive beamforming for robust multichannel speech recognition,” in *Proc. Interspeech*, San Francisco, CA, Sep 2016, pp. 1976–1980.
  - [19] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, “BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ, Dec. 2015, pp. 444–451.
  - [20] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, “Robust MVDR beamforming using time frequency masks for online offline ASR in noise,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5210–5214.
  - [21] J. Heymann, L. Drude, and R. Haeb-Umbach, “Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition,” in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016)*, San Francisco, CA, Sep 2016, pp. 12–17.
  - [22] J. Du, T. Yan-Hui, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, “The USTC-ifytek system for CHiME-4 challenge,” in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016)*, San Francisco, CA, Sep. 2016, pp. 36–38.
  - [23] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, “Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 5325–5329.
  - [24] C. Boeddeker, P. Hanebrink, L. Drude, J. Heymann, and R. Haeb-Umbach, “Optimizing neural-network supported acoustic beamforming by algorithmic differentiation,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 171–175.
  - [25] J. G. Francis, “The QR transformation — part 1,” *The Computer Journal*, vol. 4, no. 3, pp. 265–271, Jan 1961.
  - [26] S. F. Walter, L. Lehmann, and R. Lamour, “On evaluating higher-order derivatives of the QR decomposition of tall matrices with full column rank in forward and reverse mode algorithmic differentiation,” *Optimization Methods and Software*, vol. 27, no. 2, pp. 391–403, 2012.
  - [27] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. 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 *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ, Dec 2015, pp. 436–443.