PERFORMANCE OF MASK BASED STATISTICAL BEAMFORMING IN A SMART HOME SCENARIO

*Jahn Heymann**

Paderborn University, Paderborn, Germany
heymann@nt.uni-paderborn.de

Michiel Bacchiani, Tara N. Sainath

Google Inc., New York, USA
{michiel, tsainath}@google.com

ABSTRACT

Mask based statistical beamforming, where signal statistics for the target and the interference gained from masking are used for beamforming, has shown great effectiveness in the two recent CHiME challenges. This idea has sparked interest in the research community and resulted in numerous proposed approaches based on the idea. At the same time, the advent of voice controlled smart home devices, such as Google Home and Amazon Alexa, has strengthened the need for robust far-field automatic speech recognition. In this paper, we evaluate if mask based beamforming can live up to the expectations created by the CHiME challenges and provide similar gains in a smart home scenario. To this extend, we pinpoint the main differences between the scenarios, review the recent developments and conduct extensive experiments on large scale data. These experiments show that, while a $10\,\%$ relative reduction of the word error rate can be achieved, the gains are not as high as those seen in the CHiME challenge. We also show that approaches where the front-end and back-end is trained jointly do not reach the performance level of their independently trained counterparts. On the plus side, we see a $20\,\%$ relative improvement for an evaluation set with cross-talk.

***Index Terms*—** Acoustic beamforming, multi-channel ASR, noise robust ASR, smart home

1. INTRODUCTION

In the recent CHiME 4 challenge, all Top-5 systems used mask based statistical beamforming for multi-channel feature enhancement[1]. Although their exact implementations differ in the method used to estimate the time frequency (tf) masks or the beamforming criterion, the underlying idea is the same: Given a tf mask, estimate the covariance matrices for the target and noise signal and use those matrices to obtain a beamforming vector. Apart from the performances this approach demonstrated in the challenge, it has additional desirable properties: It is independent of the microphone array (i.e. its geometry and the number of microphones) and it is robust to reverberation. This sparked the interest of the research community and numerous variants have been presented since. The performance of different mask estimators, beamforming criterions as well as postfilters have been evaluated [1, 2, 3, 4]. Joint optimization of the beamformer front-end with its mask estimator and the acoustic model back-end has been considered as well [5, 6, 7].

The previous work was done on CHiME [8] which has particular properties:

1. All scenarios are recorded outside and thus the recordings have little to no reverberation. It is known that beamformers perform much better in the absence of reverberation [9].
2. Because every speaker holds the recording device (a tablet) in more or less the same way, there is a strong prior on the position of the speaker. This can be exploited for block-online processing. E.g. [10] initializes the covariance matrices with their mean on the training set.
3. The average duration of an utterance is long, especially if the $5\,\mathrm{s}$ of context allowed for the challenge is taken into account. There are also no requirements on latency and an utterance can be processed as a whole. Combined with a nearly static speaker position, this allows for a good approximation of the covariance matrices for each frequency.
4. Except for very few utterances recorded in the cafe setting, there are no interfering speakers. The different spectral patterns of speech and the occurring noise types make it easier to separate the target speaker from the background noise. Again, this helps to estimate the statistics required for the beamforming operation.
5. The data available to train the acoustic model can be considered small by today's standards. As a result, there is a strong limitation on the noise robustness of the acoustic model. Training on all six channels can only slightly mitigate this problem as the microphone recordings are still highly correlated. But the weaker the acoustic model, the higher is the impact of pre-processing steps like beamforming on the recognition performance.

In this work, we focus on a smart home scenario. Here, the situation is very different: The device can be placed freely in a room with possibly high reverberation. This also means that there is no informed prior on the initial speaker position. The queries to be recognized by the device are typically short and consist of only a few words. These queries have to be processed with a limitation on latency to provide an acceptable user experience. Operating in a home environment, the number of possible background noise types is infinite and also includes competing speakers. To cope with all these environmental influences, a large training corpus consisting of thousands of hours of speech has been created. This speech corpus is additionally augmented with noise streams taken from YouTube and reverberated using real as well as simulated room impulse responses (RIRs). This allows us to train a robust single-channel acoustic model, capable of handling noise and reverberation scenarios.

This different environment is one of the reasons conclusions about the utility of beamforming are different. While it has shown great improvements for corpora such as CHiME, so far, implicitly

*work performed at Google

[1] http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/results.html

learning fixed multi-channel filters within the structure of the acoustic model has delivered the best results on large smart home voice search corpora [11, 12]. Another reason might be that these works do not consider mask based beamforming.

This papers examines recent developments of mask based beamforming approaches in the context of a smart home scenario with its properties described above. We recap the underlying ideas and recent developments (Sec. 2). We then conduct extensive experiments with the presented approaches (Sec. 4). We focus on a large augmented voice search dataset (Sec. 3) to evaluate the utility of mask based beamforming and compare the different proposed beamforming criteria trained independently as well as jointly with the acoustic model. We also investigate the influence of the number of microphones and different mask targets. From these experiments we conclude that the achievable gains are much smaller compared to the CHiME corpora and that the models should be trained separately and not jointly (Sec. 5).

2. MASK BASED STATISTICAL BEAMFORMING

In our multi-channel scenario, the Short Time Fourier Transform (STFT) representations of D microphone signals are gathered in a vector $\mathbf{Y}_{f,t}$ where t is the frame and f the frequency bin index. Under the narrowband approximation, the vector is a superposition of a speech component $\mathbf{X}_{f,t}$ and a noise component $\mathbf{N}_{f,t}$ which are assumed to be independent:

$$\mathbf{Y}_{f,t} = \mathbf{X}_{f,t} + \mathbf{N}_{f,t}. \tag{1}$$

The beamformer now aims to remove, or suppress the noise component. This is done by filtering the observed signal with a beamforming vector $\mathbf{w}_f$:

$$\hat{S}_{f,t} = \mathbf{w}_f^{\mathsf{H}} \mathbf{Y}_{f,t}. \tag{2}$$

Depending on the definition of the beamforming criterion, $\hat{S}_{f,t}$ is either an estimate of the speech component as observed at a reference microphone or an estimate of the source speech signal.

In this work we consider two different criterions for the beamforming vector. One maximizes the expected signal-to-noise ratio (SNR) after the beamforming operation while the other minimizes the mean squared error (MSE) between the beamformer output and a reference channel.

2.1. GEV beamformer

The Generalized Eigenvalue (GEV) (or Max-SNR) beamformer aims to maximize the output SNR of the beamforming operation [13]:

$$\mathbf{w}_f^{(\mathrm{GEV})} = \arg\max_{\mathbf{w}_f} \frac{\mathbb{E}\left[\left\|\mathbf{w}_f^{\mathsf{H}} \mathbf{X}_{f,t}\right\|^2\right]}{\mathbb{E}\left[\left\|\mathbf{w}_f^{\mathsf{H}} \mathbf{N}_{f,t}\right\|^2\right]} \tag{3}$$

This equation is known as the generalized Rayleigh quotient and the solution of this optimization problem is the eigenvector corresponding to the largest eigenvalue of the generalized eigenvalue problem

$$\mathbf{\Phi}_f^{\mathbf{XX}} \mathbf{w}_f = \lambda \mathbf{\Phi}_f^{\mathbf{NN}} \mathbf{w}_f,$$

where $\mathbf{\Phi}_f^{\mathbf{XX}} = \mathbb{E}\left[\mathbf{X}_{f,t}{\mathbf{X}_{f,t}}^{\mathsf{H}}\right]$ is the Cross-Power Spectral Density (PSD) matrix of the speech signal and $\mathbf{\Phi}_f^{\mathbf{NN}} = \mathbb{E}\left[\mathbf{N}_{f,t}{\mathbf{N}_{f,t}}^{\mathsf{H}}\right]$ of the noise signal respectively.

The solution to this problem is unique up to a multiplication with a complex scalar. Since Eq. 3 is solved independently for each frequency, this can also introduce arbitrary distortions. We compute the solution by decomposing $\mathbf{\Phi}_f^{\mathbf{NN}}$ with a Cholesky decomposition, resulting in a similar regular eigenvalue problem with a Hermitian matrix. To arrive at the solution of the generalized eigenvalue problem, the resulting eigenvector is projected back with ${\mathbf{L}_f}^{-\mathsf{H}}$ where $\mathbf{L}_f \mathbf{L}_f^{\mathsf{H}} = \mathbf{\Phi}_f^{\mathbf{NN}}$. The eigenvector itself is scaled to unit norm so the scaling is determined by the noise PSD matrix. As noted in [2], this leads to a constant residual noise power for all frequencies, i.e.

$$\mathbb{E}\left[\left\|{\mathbf{w}_f^{(\mathrm{GEV})}}^{\mathsf{H}} \mathbf{N}_{f,t}\right\|^2\right] = 1, \ \forall f.$$

So in practice the scaling is not arbitrary but well defined by the noise PSD matrix. Note that this still introduces distortions but these distortions can benefit the recognition. Due to the projection of the unit norm eigenvector frequencies with high noise energy are suppressed while those with low noise energy are emphasized. Nonetheless, it can also degrade the performance since such distortions do not match the expectations of the acoustic model back-end. We therefore optionally scale the noise PSD as follows:

$$\tilde{\mathbf{\Phi}_f^{\mathbf{NN}}} = \frac{\mathbf{\Phi}_f^{\mathbf{NN}}}{\mathrm{tr}\{\mathbf{\Phi}_f^{\mathbf{NN}}\}}$$

2.2. MWF

The Multi-channel Wiener Filter (MWF) aims to minimize the distance between the beamformer output and the speech signal as received by a reference microphone when no noise is present. Assuming noise and speech are uncorrelated and introducing a trade-off factor μ, this is expressed by

$$\mathbf{w}_f^{(\mathrm{MWF})} = \min_{\mathbf{w}_f} \mathbb{E}\left[\left\|\mathbf{w}_f^{\mathsf{H}} \mathbf{X}_{f,t} - \mathbf{X}_{f,t}^{\mathrm{ref}}\right\|^2\right] + \mu\, \mathbb{E}\left[\left\|\mathbf{w}_f^{\mathsf{H}} \mathbf{N}_{f,t}\right\|^2\right].$$

The solution to this optimization problem is given by

$$\mathbf{w}_f^{(\mathrm{MWF})} = \frac{{\mathbf{\Phi}_f^{\mathbf{NN}}}^{-1} \mathbf{\Phi}_f^{\mathbf{XX}}}{\mu + \mathrm{tr}\left\{{\mathbf{\Phi}_f^{\mathbf{NN}}}^{-1} \mathbf{\Phi}_f^{\mathbf{XX}}\right\}} \mathbf{u}^{\mathrm{ref}}, \tag{4}$$

with $\mathbf{u}^{\mathrm{ref}}$ being a one-hot vector selecting the reference microphone and $\mathbf{\Phi}_f^{\mathbf{XX}} = \phi^{XX} \mathbf{a}_f \mathbf{a}_f^{\mathsf{H}}$ a Rank-1 matrix with the Acoustic Transfer Function (ATF) $\mathbf{a}_f$ [14]. The Rank-1 property can also be enforced, leading to an improved performance [2]. Depending on the choice of μ, the formulation either resembles the well-known Minimum Variance Distortionless Response (MVDR) beamformer ($\mu \to 0$), the minimum MSE solution ($\mu = 1$) or a trade-off. It is also possible to choose a frequency dependent value for μ. Especially setting $\mu_f = \sqrt{\phi_f^{XX} \rho_f} - \rho_f$ with $\rho_f = \mathrm{tr}\left\{{\mathbf{\Phi}_f^{\mathbf{NN}}}^{-1} \mathbf{\Phi}_f^{\mathbf{XX}}\right\}$ yields the same constant value for the residual noise power along frequencies as the GEV beamformer [2].

Table 1: Network configuration for mask estimation

	Units	Type	Non-Linearity	p_{dropout}
L1	512	(B)LSTM	Tanh	0.5
L2	1024	FF	ReLU	0.5
L3	1024	FF	ReLU	0.5
L4	257 + 257	FF	Sigmoid	0.0

2.3. Mask estimator

The beamformer variants described above rely on the knowledge of the frequency-dependent PSD matrices of the speech and noise signal respectively. These are estimated from the observed signal with a mask indicating for each tf-bin if the speech or the noise is predominant:

$$\mathbf{\Phi}_f^{\nu\nu} = \frac{1}{Z} \sum_t M_{f,t}^{\nu} \mathbf{Y}_{f,t} \mathbf{Y}_{f,t}^{\mathsf{H}} \tag{5}$$

Here, Z is a normalization constant, e.g. $\sum_t M_{f,t}^{\nu}$ and ν is either $\mathbf{X}$ or $\mathbf{N}$.

The masks are estimated with a neural network separately for speech and noise. In this work, we employ two different network architectures (see Table 1): One bi-directional network and one operating frame-by-frame. The former is inspired by the architecture used in [15]. The later replaces the bi-directional Long Short-Term Memory (BLSTM) with an LSTM and also uses a different kind of normalization. While the BLSTM network normalizes, scales and shifts after each linear transformation (i.e. before applying the non-linearity), the LSTM network only normalizes the mean after the linear transformation of the LSTM layer using cumulative statistics instead of utterance level statistics.

Both networks are trained with a cross-entropy criterion. Using simulated data (see Sec. 3), we have oracle information to calculate target masks for the training. For those, we consider three different options:

1. As in [15] we compute ideal-binary masks with different thresholds for the speech ($10\,\mathrm{dB}$) and noise ($-5\,\mathrm{dB}$). The masks do not sum to one here and the speech mask is sparse as we only want to include tf-bins that clearly are dominated by speech in the calculation of the PSD matrices. The resulting masks are binary.
2. Ratio masks which have shown to perform well and can also be used as a smooth postfilter [3].
3. We only consider those tf-bins as dominated by speech, where the instantaneous power is above the average power of the signal. This, again, yields a very sparse speech mask.

The mask estimator operates on each microphone channel individually and the masks are pooled with a median operation resulting in a single mask for speech and noise to be used in Eq. 5. To avoid a transformation back to the time-domain the mask estimator as well as the beamformer operates in the spectral domain with a frame size of $25\,\mathrm{ms}$ and a frame shift of $10\,\mathrm{ms}$.

2.4. Implementation

All beamformer models have been implemented in Tensorflow [16] with support for backpropagation through the respective optimization criterion [17]. This allowed us to combine the front-end (beamformer) with the back-end (acoustic model) into one model with a complex-valued multi-channel input and integrated statistical beamforming. The resulting model can be trained end-to-end (E2E) with any speech/phoneme classification criterion [7].

However, the training of such a model can quickly become unstable. Especially at the beginning of the training, it might happen that only a few tf-bins are considered as noise for some frequencies. Then, due to numerical issues, the noise PSD might not be positive definite as required for the Cholesky decomposition. We therefore enforce this property by decomposing the noise PSD matrix and shifting its eigenvalues to the positive regime if required. This shift is ignored during the backpropagation pass.

2.5. Acoustic model

The acoustic model consists of 5 Long Short-Term Memory (LSTM) layers with 768 units per layer. It is trained to classify the 8,192 context-dependent phonemes [18] with a cross-entropy criterion without any further sequential fine-tuning. We stack three consecutive frames so the model operates at a reduced framerate of $33\,\mathrm{Hz}$ and also delay the targets for $150\,\mathrm{ms}$ [19]. This model has proven to be a strong baseline internally.

3. DATA

Our training data set consists of $250\,\mathrm{k}$ voice search query utterances translating into roughly 150 hours of speech. The training set is anonymized and hand-transcribed, and is representative of Googles voice search traffic. The data is corrupted using a room simulator adding varying degrees of noise and reverberation. The speaker is placed in one of 100 sampled room configurations with T_{60} times ranging from $400\,\mathrm{ms}$ to $900\,\mathrm{ms}$ with an average of roughly $600\,\mathrm{ms}$. The distance between the speaker and the circular array with 8 microphones ranges from $1\,\mathrm{m}$ to $4\,\mathrm{m}$. The radius of the circular array is $7\,\mathrm{cm}$. Noise is added by placing 0 to 3 noise source into the same room. The noise signals are extracted from YouTube and include samples with music and ambient noise. The final SNR ranges from $0\,\mathrm{dB}$ to $20\,\mathrm{dB}$ with an average of about $12\,\mathrm{dB}$.

For evaluation we use $100\,\mathrm{k}$ utterances and corrupt them in the same way as described above. The characteristics are sampled from the same distributions but we ensure that the samples do not match with one from the training set. We also create a second evaluation set that uses speech samples from YouTube as noise source to simulate a multi-talker environment.

4. EXPERIMENTAL RESULTS

The word error rate (WER) results for the different beamformer criterions are shown in Tbl. 2. Both criterions are able to improve the recognition performance over the single channel baseline by about $10\,\%$ relative. This is already the case if only 2 channels are used (microphone 1 and 5 of the circular array), especially for the MWF. Adding two more microphones leads to an improvement for the GEV but the MWF does not profit from the additional channel. Increasing the number to eight brings no further gains for the GEV and the performance of the MWF degrades. The reason for this is probably the close microphone spacing [20]. Overall, the difference between the GEV and the MWF is marginal and depends on the number of microphones used as well as their respective parameters/normalization. Note that the table only shows the results for the best configuration.

Table 2: Comparison between GEV and MWF for different numbers of microphones with and without cross-talk (CT). For each scenario the parameters yielding the lowest WER have been chosen.

	1 ch	2 ch	4 ch	8 ch
GEV	30.6	28.4	27.3	27.4
MWF		27.7	27.7	29.3
GEV CT	34.8	29.6	29.1	28.6
MWF CT		29.6	29.6	32.1

Table 3: Comparison of mask targets averaged for GEV and MWF.

	IBM	RATIO	AVG
noisy	27.8	29.3	27.7
CT	30.0	31.5	29.6

In general, we found that for the GEV the best results are achieved without any normalization applied to the noise PSD or the beamforming vector itself. For the MWF, we found that it is beneficial to approximate the speech PSD matrix with a Rank-1 matrix but without considering the noise matrix (see Sec 2.2 and [2]). The generalized decomposition leads to unstable behaviour for some utterances, resulting in overall worse performance. No clear statement can be made about the choice of the trade-off parameter. Sometimes the frequency dependent one worked better, sometimes setting it to 0 (MVDR) yielded better results. There was also no noticeable difference between the two mask estimators.

Further investigations shows that most gains are achieved for low SNR and/or high T_{60} conditions. Especially the good performance for high reverberation was not expected since the narrowband assumptions is violated in those cases and beamformers are said to perform worse in those conditions.

For the evaluation set with cross-talk (CT), the gains achieved by using a multi-channel input are larger, resulting in a relative reduction of about 20 %. The MWF shows the same behaviour as described above regarding the number of microphones. Using two or four makes no difference but changing to eight deteriorates the recognition. The GEV can profit from an increasing number of microphones and achieves slightly better results overall in this scenario. Neither the mask estimator, nor the acoustic model are trained on cross-talk. However, if the mask estimator confuses some tf-bins it does not hurt the estimation of the PSD matrices too much as long as the majority is correct for a given frequency. In contrast, the acoustic model is more sensitive to interferences from another speaker.

The results for the different mask targets are shown in Tbl. 3. From those, we conclude that the choice of the target is not very important. One tendency that can be seen is that it is beneficial to use sparse masks, i.e. only relying on a few but reliable tf-bins yields a better estimate for the PSD matrices. Consequently, the ratio mask performs slightly worse compared to the other two.

The results for the E2E training are shown in Tbl. 4. Concentrating on the results for *E2E GEV* & *E2E GEV CT* first, we can make two observations: First, the results are worse compared to the one with independent models and even compared to the baseline. Second, the results now clearly depend on the number of microphones. Regarding the first point, our initial guess was that the unreliable acoustic model at the beginning of the training might lead to incorrect gradients for the mask estimator pushing it into a direction it can never recover from. As a consequence, we additionally used the cross-entropy mask loss (with IBM masks) as an auxiliary loss (+mask). This lead to a significant performance improvement but we were unable to match the WERs achieved with separate models even though the masks produced by the model showed the same quality in terms of cross-entropy loss. As a result of the second observation, we started to sample the number of microphones uniformly during the training (+sampling). As an interesting result, the model now achieves the best performance using four microphones. From these experiments we conclude that training the model jointly does not yield the same (or even better) performance than training each component individually. While our initial guess in [7] was that this might be a matter of available data, we can reject this hypothesis. Instead, we hypothesize that, since the masks and thus the signal enhancement were comparable, training both components jointly leads to a weaker acoustic model and an overfitting to the specific characteristics of the beamformer output, i.e. the same effect observed when training the acoustic model on beamformed features [21].

Table 4: Results for the E2E approaches.

	2 ch	4 ch	8 ch
E2E GEV	42.1	38.6	37.8
E2E GEV + mask	37.3	31.8	30.4
E2E MWF + mask	40.3	35.0	33.0
E2E GEV + mask + sampling	39.4	34.9	36.1
E2E GEV CT	50.2	44.8	41.0
E2E GEV + mask CT	41.6	34.6	33.1
E2E MWF + mask CT	45.8	39.9	41.3
E2E GEV + mask + sampling CT	45.9	39.9	37.5

5. CONCLUSIONS

This paper evaluates various mask based beamforming variants for a smart home scenario on large scale data. The results show that we can achieve a 10 % relative gain over a single-channel baseline by using just two microphones. The gain is even larger when cross-talk is considered as a possible noise source (20 % relative). Increasing the number of microphones further yields only small gains with the chosen geometry. We also evaluated E2E approaches but found that those cannot reach the performance of their independently trained counterparts, presumably mainly due to a weaker acoustic model. Taking into account the flexibility independently trained models provide and the difficulties arising during the training of the E2E models, it is hard to find an advantage for the latter. Overall, we can conclude that the mask based beamformer can provide gains for the scenario in question. However, these are not as high as the ones seen on the CHiME corpora. We think the main reason for this is that an acoustic model trained on 150 hours of data is more robust than one trained on roughly 18 hours. One thing we did not take into account is the latency constraint. While both, the (LSTM) mask estimator used as well as the acoustic model can operate on a frame-by-frame basis, the statistics are still accumulated over the whole utterance. Initial experiments on block-wise processing showed that it is hard to maintain the gains achieved here. Optimizing this is an open question for future research.

6. REFERENCES

[1] Hakan Erdogan, John Hershey, Shinji Watanabe, Michael Mandel, and Jonathan Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep 2016.

[2] Ziteng Wang, Emmanuel Vincent, Romain Serizel, and Yonghong Yan, "Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments," Jul 2017.

[3] Xueliang Zhang, Zhong Qiu Wang, and DeLiang Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2017.

[4] Lukas Pfeifenberger, Matthias Zohrer, and Franz Pernkopf, "DNN-based speech mask estimation for eigenvector beamforming," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2017.

[5] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R. Hershey, and Xiong Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, Dec 2017.

[6] Xiong Xiao, Shengkui Zhao, Douglas L. Jones, Eng Siong Chng, and Haizhou Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2017.

[7] Jahn Heymann, Lukas Drude, Christoph Boeddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2017.

[8] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, no. C, Nov. 2017.

[9] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka, "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.

[10] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2016.

[11] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani, and Andrew, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015.

[12] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, and Michiel Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2016.

[13] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming based on Generalized Eigenvalue Decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, 2007.

[14] Mehrez Souden, Jacob Benesty, and Sofine Affes, "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, Feb 2010.

[15] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2016.

[16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[17] Christoph Boeddeker, Patrick Hanebrink, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "On the Computation of Complex-valued Gradients with Application to Statistically Optimum Beamforming," *arXiv:1701.00392 [cs.NA]*, 2017.

[18] Andrew Senior, Hasim Sak, Felix de Chaumont Quitry, Tara N. Sainath, and Kanishka Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[19] Golan Pundak and Tara N. Sainath, "Lower frame rate neural network acoustic models," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep 2016.

[20] Gary W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 61–85. Springer, 2001.

[21] Emmanuel Vincent, "When mismatched training data outperform matched data," in *Systematic approaches to deep learning methods for audio*, 2017.