INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Eigenvector-based Speech Mask Estimation using Logistic Regression

Lukas Pfeifenberger[1,2], *Matthias Zöhrer*[1], *Franz Pernkopf*[1]

[1] Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
[2] Ognios GmbH Salzburg, Austria

lukas.pfeifenberger@alumni.tugraz.at
{matthias.zoehrer,pernkopf}@tugraz.at

Abstract

In this paper, we use a logistic regression to learn a speech mask from the dominant eigenvector of the *Power Spectral Density* (PSD) matrix of a multi-channel speech signal corrupted by ambient noise. We employ this speech mask to construct the *Generalized Eigenvalue* (GEV) beamformer and a Wiener postfilter. Further, we extend the beamformer to compensate for speech distortions. We do not make any assumptions about the array geometry or the characteristics of the speech and noise sources. Those parameters are learned from training data. Our assumptions are that the speaker may move slowly in the near-field of the array, and that the noise is in the far-field. We compare our speech enhancement system against recent contributions using the CHiME4 corpus. We show that our approach yields superior results, both in terms of perceptual speech quality and speech mask estimation error.

Index Terms: Multi-channel speech enhancement, broadband beamforming, speech mask estimation

1. Introduction

In many beamforming structures, a *steering vector* is required to provide a spatial focus towards the location of the speaker. A simple and robust method is to obtain the steering vector from a *Direction Of Arrival* (DOA) estimate. Many algorithms have been devised for that purpose, i.e. PHAT, MUSIC [1], or DD-SNR [2]. However, the DOA cannot model reverberation or multi-path propagation caused by the enclosure, i.e. office rooms or car interiors. This may result in target leakage, which limits beamforming performance. More advanced approaches aim at estimating the acoustic path from the speech source to each microphone, which is known as *Acoustic Transfer Function* (ATF). Approximations are done using *Relative Transfer Functions* (RTFs) [3,4]. The RTFs relate the ATFs with respect to a reference point, and can be modeled by shorter FIR filters [5]. Recent contributions use a spectral gain mask to distinguish between speech and noise signals, which is then used to estimate their respective PSD matrices. Such a *speech mask* may be obtained using model-based clustering approaches [6–8], or data-driven regression [9–12] based on various types of *neural networks* (NN). While clustering approaches require some prior knowledge like the array geometry or the statistics of the noise, NNs are able to learn the speech mask from training data, without additional information. Moreover, NNs have the distinct advantage of jointly estimating a speech mask for all frequencies, which proved to be superior in recent multi-channel speech enhancement and recognition tasks [13, 14].

In this paper, we extend our work in [12], where we used several NN architectures to estimate the speech mask using eigenvector features. As the largest NN requires over a million weights to be trained, the aim is to significantly reduce model complexity, while maintaining performance. We introduce a different approach to estimate the speech mask compared to [8] and [9]. Instead of energy-related features, our NN utilizes the dominant eigenvector of the PSD matrix of the microphone signals as feature vector. Therefore, the spatial information hidden in the multi-channel data is exploited. The predicted speech mask is then used to split the PSD matrix of the microphone signals into its speech and noise components, where we use the dominant eigenvector of the speech PSD as steering vector for the beamformer. We show that the cosine similarity between dominant eigenvectors of consecutive PSD matrices of the microphone signals is sufficient to predict the speech mask. By using the cosine similarity, we obtain a feature which is independent of both the signal energy and the microphone array geometry. Our assumptions are that the speaker is in the near-field of the array and that the non-stationary noise is in the far-field. The speaker may move slowly, resulting in slowly varying ATFs. These relaxed conditions are found in many telephony applications, i.e. hands-free calling kits, voice chats on mobile devices, or roadside emergency telephones.

This paper is structured as follows: After the introduction of the system model in Section 2, we demonstrate our extension to the GEV beamformer for reducing speech distortions in Section 3. In Section 4 we show the importance of the speech presence probability for constructing the beamformer and a postfilter. In Section 5 the estimation of the speech presence probability using logistic regression is presented. In Section 6 we present our results. Section 7 concludes the paper.

2. System Model

We assume a single speech source embedded in ambient noise. The array consists of M microphones, arranged into an arbitrary array geometry. There may be multiple noise sources, and their spatial and temporal characteristics may be unknown. Our speech enhancement system is shown in Figure 1. We define

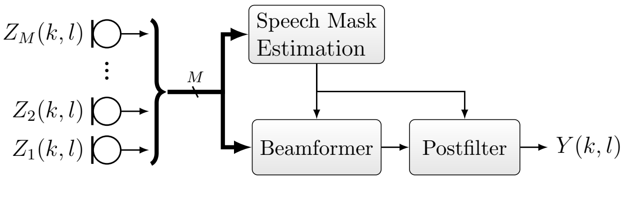

Figure 1: *System overview, showing the microphone signals $Z_m(k,l)$ and the beamformer+postfilter output $Y(k,l)$ in frequency domain.*

http://dx.doi.org/10.21437/Interspeech.2017-1186

the signal at the m^{th} microphone in the STFT domain as

$$Z_m(k,l) = S(k,l)A_m(k,l) + N_m(k,l), \tag{1}$$

where the frequency bin $k = 1, \ldots, K$ and the time frame is denoted by l. The ATF of the speaker $S(k,l)$ to the m^{th} microphone is denoted by $A_m(k,l)$, and the noise received at that microphone is denoted by $N_m(k,l)$. By stacking all M signals to a $M \times 1$ vector, the signal model can be written as:

$$\boldsymbol{Z}(k,l) = S(k,l)\boldsymbol{A}(k,l) + \boldsymbol{N}(k,l). \tag{2}$$

For enhanced readability, the frequency and time frame indices will be omitted except where necessary. The PSD matrix for all microphone signals $\boldsymbol{Z}(k,l)$ is obtained using recursive averaging, i.e.: $\boldsymbol{\Phi}_{ZZ}(k,l) = \boldsymbol{\Phi}_{ZZ}(k,l-1)\alpha + (1-\alpha)\boldsymbol{Z}(k,l)\boldsymbol{Z}^H(k,l)$, where $0 \leq \alpha \leq 1$ is a smoothing parameter [15]. For uncorrelated speech and noise signals, this PSD matrix can be split into its speech and noise components

$$\boldsymbol{\Phi}_{ZZ}(k,l) = \boldsymbol{\Phi}_{SS}(k,l) + \boldsymbol{\Phi}_{NN}(k,l). \tag{3}$$

For a single speaker, $\boldsymbol{\Phi}_{SS}(k,l)$ will be of rank 1 and can therefore be decomposed into

$$\boldsymbol{\Phi}_{SS}(k,l) = \boldsymbol{A}(k,l)\boldsymbol{A}^H(k,l)\Phi_S(k,l). \tag{4}$$

Note that the magnitude of the ATFs can be modeled by $\Phi_S(k,l)$ in (4), hence we define $||\boldsymbol{A}||^2 = 1$.

3. Multi-Channel Speech Enhancement

We use a *broadband* beamformer [16] for multi-channel speech enhancement. The beamformer output is given by

$$Y(k,l) = \boldsymbol{W}^H(k,l)\boldsymbol{Z}(k,l)G(k,l), \tag{5}$$

with the filter weights $\boldsymbol{W}(k,l)$, and the single-channel Wiener postfilter $G(k,l)$. Following the definition of the system model in (2), the optimal filter weights are given by the MSE-optimal multi-channel Wiener filter $\boldsymbol{W}_{OPT}$ [15, 17], i.e.

$$\boldsymbol{W}_{OPT} = \underbrace{\frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}}_{\boldsymbol{W}_{MVDR}} \cdot \underbrace{\frac{\Phi_S}{\Phi_S + \left[\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}\right]^{-1}}}_{G=\frac{\xi}{1+\xi}}. \tag{6}$$

The filter $\boldsymbol{W}_{MVDR}$ can be recognized as the MVDR beamformer [18, 19]. The postfilter $G = \frac{\xi}{1+\xi}$ depicts a real-valued gain mask, which is applied at the beamformer output. It can be seen from (4) and (6), that ξ is given as the multi-channel SNR [17]

$$\xi = \Phi_S\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A} = \mathrm{Tr}\{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{\Phi}_{SS}\}. \tag{7}$$

3.1. GEV Beamformer

While it is possible to select from a broad range of broadband beamformers such as the MVDR or the GSC, we use the GEV for its superior performance in earlier experiments [12]. The GEV beamformer, constrains the filter weights $\boldsymbol{W}(k,l)$ to maximize the SNR $\xi(k,l)$ at the beamformer output [20, 21], i.e.

$$\boldsymbol{W}_{GEV} = \arg\max_{\boldsymbol{W}} \xi. \tag{8}$$

The solution to (8) is given by the following generalized eigenvalue problem

$$\boldsymbol{\Phi}_{SS}\boldsymbol{W}_{GEV} = \xi\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{GEV}, \tag{9}$$

which is solved by

$$\boldsymbol{W}_{GEV} = \zeta\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}, \tag{10}$$

where ζ is an arbitrary complex scalar. Comparing the GEV to the MVDR beamformer, it can be immediately seen that they only differ by a complex constant C

$$\boldsymbol{W}_{MVDR} = \frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}} = C\boldsymbol{W}_{GEV} \tag{11}$$

However, this difference causes target speech distortions in the GEV, i.e. $\boldsymbol{W}_{GEV}^H(k,l)\boldsymbol{A}(k,l) \neq 1$. To compensate for these distortions, we derive an expression for C as follows:

Assuming normalized ATFs $||\boldsymbol{A}||^2 = 1$, we can rearrange (10) into $\zeta = \boldsymbol{A}^H\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{GEV}$ and express the complex constant C by

$$C_{PAN} = \frac{\boldsymbol{W}_{GEV}^H\boldsymbol{\Phi}_{NN}\boldsymbol{A}}{\boldsymbol{W}_{GEV}^H\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{GEV}}, \tag{12}$$

which we refer to as *Phase Aware Normalization* (PAN). Note that [20] proposes the *Blind Analytical Normalization* (BAN) and the *Blind Statistical Normalization* (BSN) compensation methods to estimate the absolute value of C, i.e.: $C_{BAN} = \frac{\sqrt{\boldsymbol{W}_{GEV}^H\boldsymbol{\Phi}_{NN}\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{GEV}}}{\boldsymbol{W}_{GEV}^H\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{GEV}}$. In fact, it can be easily verified that the magnitudes of the BAN and PAN compensation factors are identical. Inserting (12) into (11) gives the GEV-PAN beamformer:

$$C_{PAN}\boldsymbol{W}_{GEV} = \frac{\boldsymbol{W}_{GEV}\boldsymbol{W}_{GEV}^H\boldsymbol{\Phi}_{NN}}{\boldsymbol{W}_{GEV}^H\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{GEV}}\boldsymbol{A} = \boldsymbol{P}\boldsymbol{A} \tag{13}$$

with the projection matrix $\boldsymbol{P}$ [22]. The expression $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{P}$ can be identified as blocking matrix [21]. In theory, the compensation factor C_{PAN} turns the GEV into the MVDR beamformer. However, as the former avoids the inversion of $\boldsymbol{\Phi}_{NN}$ when using (9), it is numerically more stable and achieves better PESQ and OPS scores [12].

3.2. Steering Vector Estimation

From (13) it can be seen that the GEV-PAN beamformer requires the ATFs $\boldsymbol{A}(k,l)$. As they are unknown in practice and hard to estimate in reverberant environments [18], we use a steering vector $\boldsymbol{F}(k,l)$, which provides a spatial focus of the speech source. Under reverberation-free conditions the steering vector may be modeled as simple time delays using DOA estimation [1]. However, in realistic environments this approach will result in speech loss at the beamformer output. We therefore advocate a steering vector in signal subspace [23]. Eigenvalue decomposition (EVD) of the speech PSD matrix gives

$$\boldsymbol{\Phi}_{SS} = \boldsymbol{A}\boldsymbol{A}^H\Phi_S = \boldsymbol{v}_{S_1}\boldsymbol{v}_{S_1}^H\lambda_{S_1}, \tag{14}$$

where λ_{S_1} and $\boldsymbol{v}_{S_1}$ are the single eigenvalue and eigenvector of $\boldsymbol{\Phi}_{SS}(k,l)$, as this matrix is of rank 1 for a single speaker. Rearranging (14) leads to:

$$\boldsymbol{A} = \frac{\lambda_{S_1}}{\boldsymbol{A}^H\boldsymbol{v}_{S_1}\Phi_S}\boldsymbol{v}_{S_1} = \zeta_{S_1}\boldsymbol{v}_{S_1}, \tag{15}$$

where ζ_{S_1} can be recognized as another complex scalar. Therefore, the dominant eigenvector of $\boldsymbol{\Phi}_{SS}$ is a scaled version of the true ATF $\boldsymbol{A}(k,l)$, including multi-path propagations and early echoes of the target signal [1, 20, 23]. Assuming $||\boldsymbol{A}||^2 = 1$, the dominant eigenvector $\boldsymbol{v}_{S_1}$ is equal to the ATF.

4. Speech Mask Estimation

In the last section, we have shown that the GEV-PAN beamformer and the steering vector require an estimate of both the speech and noise PSD matrices.

4.1. PSD matrix approximation

By using an oracle speech mask $0 \leq p_{\text{SPP}}(k,l) \leq 1$, which represents the probability for each time-frequency bin to contain speech, $\mathbf{\Phi}_{SS}(k,l)$ can be approximated with

$$\hat{\mathbf{\Phi}}_{SS}(k,l) = \frac{\sum_{t=l-T/2}^{l+T/2} \mathbf{Z}(k,t)\mathbf{Z}^H(k,t)p_{\text{SPP}}(k,t)}{\sum_{t=l-T/2}^{l+T/2} p_{\text{SPP}}(k,t)}, \tag{16}$$

where T is a number of frames during which we assume the spatial characteristics of $\mathbf{\Phi}_{SS}(k,l)$ to be stationary, i.e. the speaker is not moving [24]. Note that the energy of the speech signal may change during the time period T, but this does not affect (9), and hence the performance of the GEV beamformer. Analogously to (16), the noise PSD matrix $\mathbf{\Phi}_{NN}(k,l)$ can be approximated using the complementary probability $1 - p_{\text{SPP}}(k,t)$.

4.2. Speech Presence Probability

As shown above, the estimation of $p_{\text{SPP}}(k,l)$ is the key component in our speech enhancement system. Using (7), we define the optimal speech presence probability as

$$p_{\text{SPP,opt}}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} = G(k,l). \tag{17}$$

Note that the optimal speech presence probability is equal to the Wiener postfilter given in (6).

Eigenvalue decomposition of the noisy speech PSD matrix gives $\mathbf{\Phi}_{ZZ} = \sum_{m=1}^{M} \lambda_{Z_m} \boldsymbol{v}_{Z_m} \boldsymbol{v}_{Z_m}^H$, where λ_{Z_m} and $\boldsymbol{v}_{Z_m}$ are its eigenvalues and eigenvectors, respectively. We observed that the dominant eigenvector $\boldsymbol{v}_{Z_1}(k,l)$ is related to $p_{SPP,\text{opt}}(k,l)$ [24]. It was shown that eigenvectors containing speech tend to form local clusters, while noisy eigenvectors are distributed randomly over a multi-dimensional unit sphere. Hence, the dominant eigenvector $\boldsymbol{v}_{Z_1}(k,l)$ provides a reliable measure for speaker activity. Using the cosine similarity[1] between two neighboring eigenvectors

$$x_{\Delta}(k,l) = |\boldsymbol{v}_{Z_1}(k,l)^H \boldsymbol{v}_{Z_1}(k,l-\Delta)|, \tag{18}$$

we obtain a scalar $x_{\Delta}(k,l)$, independent of the number of microphones being used. To observe a difference between two neighboring eigenvectors, the matrix $\mathbf{\Phi}_{ZZ}$ has to be updated with a sufficiently small time constant. During speaker activity, $x_{\Delta}(k,l)$ is close to one, and close to zero otherwise. Note that this feature is also independent of signal energy and array geometry. Figure 2 shows $x_{\Delta}(k,l)$ for a single utterance of the CHiME4 corpus. It can be seen that $x_{\Delta}(k,l)$ already has some similarity with the optimal speech mask $p_{\text{SPP,opt}}(k,l)$, shown in Figure 3a. At low frequencies, the separation capability of this feature is poor, as the wavelength of the signal is large compared to the aperture of a typical microphone array. This information has to be inferred from other frequency components.

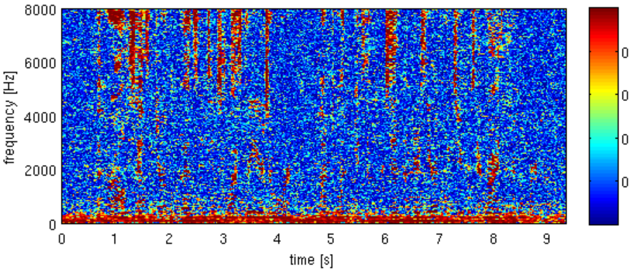

Figure 2: *Cosine similarity $x_{\Delta=1}(k,l)$ for a single utterance of the CHiME4 corpus.*

5. Logistic Regression

In contrast to the NNs based on LSTMs and MLPs used in [12], we aim for a resource-efficient regression model to estimate the speech mask. As our model operates in the time-frequency domain, we derive a feature vector for each frequency bin k and time frame l. In particular, we stack n_{Δ} cosine distances $x_{\Delta}(k,l)$ to add some context to the feature vector

$$\boldsymbol{x}_{\text{evd}}(k,l) = \left[x_{\Delta=1}(k,l), \ldots, x_{\Delta=n_{\Delta}}(k,l)\right]^T, \tag{19}$$

where we consider the eigenvectors in the vicinity n_{Δ} of the current time frame l containing the most relevant information. We refer to (19) as *eigenvector-delta*. For each frequency bin k we obtain an estimate

$$\tilde{p}_{\text{SPP}}(k,l) = \frac{e^{\tilde{a}(k,1)}}{\sum_{i=1}^{2} e^{\tilde{a}(k,i)}}, \tag{20}$$

using the activation

$$\tilde{a}(k,i) = \tilde{\boldsymbol{W}}(k,i)\boldsymbol{x}_{\text{evd}}(k,l) + \tilde{b}(k,i), \tag{21}$$

where $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{b}}$ denote the weights and bias values of the logistic regression, respectively. Note that $\tilde{p}_{\text{SPP}}(k,l)$ is calculated independently for each frequency bin. To exploit the broadband nature of human speech, we employ a second logistic regression which calculates a refined estimate $\hat{p}_{\text{SPP}}(k,l) = \frac{e^{\hat{a}(k,1)}}{\sum_{i=1}^{2} e^{\hat{a}(k,i)}}$. The activation $\hat{a}(k,i)$ uses the estimate $\tilde{p}_{\text{SPP}}(k,l)$ of the neighboring $k - k_{\Delta} \cdots k + k_{\Delta}$ frequency bins, i.e.

$$\hat{a}(k,i) = \sum_{j=k-k_{\Delta}}^{k+k_{\Delta}} \hat{\boldsymbol{W}}(k,j,i)\tilde{p}_{\text{SPP}}(j,l) + \hat{b}(k,i). \tag{22}$$

This architecture is capable to learn the basic structure of human speech. As the eigenvector-features do not contain any information about signal energy, *speaker-dependent* features are ignored.

6. Results

6.1. Experimental Setup

To evaluate our model, we use the CHiME4 corpus [13], which provides 2 and 6-channel recordings of a close-talking speaker corrupted by four different types of ambient noise. The database provides a training set (tr05), a validation set (dt05) and a test set (et05). We use all utterances (real and simu) from each set. The ground truth (i.e. the separated speech and noise signals) is available for all recordings, which we use to calculate the true speech mask $p_{\text{SPP,opt}}(k,l)$ with (7) and (17). Once

[1] Note that the eigenvectors are already normalized to 1, i.e. $||\boldsymbol{v}_{Z_1}(k,l)||_2^2 = 1$.

trained, the logistic regression provides a prediction $\hat{p}_{\text{SPP}}(k,l)$ for each utterance, which is required to calculate $\hat{\mathbf{\Phi}}_{SS}(k,l)$ and $\hat{\mathbf{\Phi}}_{NN}(k,l)$ with (16). The averaging window length is set to $T = 250ms$. We use a STFT window length of 32ms and an overlap of 50% to process the data. The speech and noise PSD estimates are then used to construct the GEV-PAN beamformer. Following (17), we use $G(k,l) = \hat{p}_{\text{SPP}}(k,l)$ for the postfilter.

6.2. Speech Mask Accuracy

Figure 3 shows the performance of the logistic regression models by visualizing the optimal and predicted speech masks for a single utterance from the test set. Table 1 reports the prediction error $\mathcal{L} = \frac{100}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\left|\hat{p}_{\text{SPP}}(k,l) - p_{\text{SPP,opt}}(k,l)\right|$ in % for the logistic regression, labeled as evd_logreg, and four alternative NN models which we used in [12]. They are labeled as ev_lstm, evd_lstm, evd_mlp and psd_lstm, and use multiple hidden layers, each containing n_h neurons. The last column shows the number of parameters to be trained for each model. It can be seen that the logistic regression uses two orders of magnitude fewer weights while achieving comparable results.

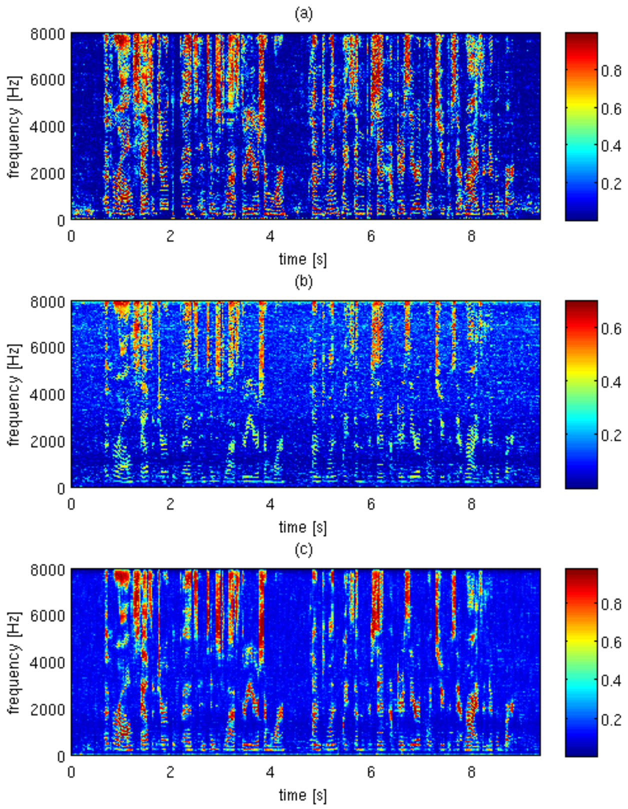

Figure 3: *Speech presence probability for a single utterance from the CHiME4 test set. (a) ground truth $p_{\text{SPP,opt}}(k,l)$, (b) coarse prediction $\tilde{p}_{\text{SPP}}(k,l)$, (c) refined prediction $\hat{p}_{\text{SPP}}(k,l)$.*

Table 1: *Prediction error for $\hat{p}_{SPP}(k,l)$ in % for the 6 channel data. Results of proposed methods are bold face.*

architecture	n_Δ	n_h	*train*	*valid*	*test*	*# of weights*
ev_lstm, 6ch	-	20,10	1.889	2.685	3.003	1457704
evd_lstm, 6ch	7	20,10	2.184	2.183	2.520	1252104
evd_mlp, 6ch	7	20,10	2.349	2.285	2.825	156513
psd_lstm, 6ch	-	20,10	2.711	3.415	3.489	1210984
evd_logreg, $\tilde{p}_{\text{SPP}}$, 6ch	3	-	**3.980**	**3.894**	**5.700**	**1542**
evd_logreg, $\hat{p}_{\text{SPP}}$, 6ch	3	-	**2.767**	**2.671**	**3.598**	**12336**
ev_lstm, 2ch	-	10,5	3.919	4.265	4.377	1128744
evd_lstm, 2ch	7	10,5	3.566	3.495	3.992	1252104
evd_mlp, 2ch	7	10,5	3.695	3.613	4.778	156513
psd_lstm, 2ch	-	10,5	4.620	5.082	4.902	1046504
evd_logreg, $\tilde{p}_{\text{SPP}}$, 2ch	3	-	**6.575**	**6.493**	**7.497**	**1542**
evd_logreg, $\hat{p}_{\text{SPP}}$, 2ch	3	-	**4.382**	**4.282**	**5.901**	**13364**

6.3. Perceptual Speech Quality

Using the predicted speech mask $\hat{p}_{\text{SPP}}(k,l)$, we construct the GEV-PAN beamformer for both the 2 and 6-channel data. To evaluate the performance of the resulting speech signal $Y(k,l)$ in terms of perceptual speech quality, we use the *Perceptual Evaluation methods for Audio Source Separation* (PEASS) toolkit [25], and report the *Overall Perceptual Score* (OPS) and PESQ [26] values. The ground truth required for these scores is obtained using the $p_{\text{SPP,opt}}(k,l)$ and the GEV-PAN. Table 2 reports the PESQ and OPS scores of the logistic regression and the other models used in Table 1. Further, we also report the scores of the CHiME4-baseline enhancement system, i.e. the BeamformIt!-toolkit [13], and the front-end of the best CHiME3 system [8], which uses CGMM priors and the EM algorithm to estimate the speech mask. It can be seen that the performance of the much smaller logistic regression architecture (evd_logreg) is comparable to the NN models, even for the 2-channel track. For this track, 2 out of the 6 microphones have been chosen randomly, so that the array geometry is unknown for each utterance. In summary, all our eigenvector-based speech mask estimation models show an improvement over the BeamformIt! baseline and the CGMM-EM system.

Table 2: *Performance comparison of 6- and 2-channel data, against the BeamformIt! and CGMM-EM systems.*

architecture	n_Δ	n_h	*PESQ [MOS]*			*OPS [%]*		
			train	*valid*	*test*	*train*	*valid*	*test*
ev_lstm, 6ch	-	20,10	2.443	2.007	1.891	72	58	51
evd_lstm, 6ch	7	20,10	2.226	1.969	1.874	67	59	52
evd_mlp, 6ch	7	20,10	2.197	1.944	1.829	67	59	52
psd_lstm, 6ch	-	20,10	1.977	1.758	1.724	63	54	49
evd_logreg, $\tilde{p}_{\text{SPP}}$, 6ch	3	-	**1.830**	**1.676**	**1.551**	**57**	**52**	**46**
evd_logreg, $\hat{p}_{\text{SPP}}$, 6ch	3	-	**2.071**	**1.862**	**1.704**	**63**	**57**	**50**
ev_lstm, 2ch	-	10,5	1.965	1.706	1.725	51	44	45
evd_lstm, 2ch	7	10,5	2.090	1.850	1.869	48	43	43
evd_mlp, 2ch	7	10,5	2.042	1.818	1.820	46	42	42
psd_lstm, 2ch	-	10,5	1.867	1.669	1.703	44	40	41
evd_logreg, $\tilde{p}_{\text{SPP}}$, 2ch	3	-	**1.696**	**1.578**	**1.579**	**35**	**32**	**34**
evd_logreg, $\hat{p}_{\text{SPP}}$, 2ch	3	-	**1.940**	**1.754**	**1.741**	**43**	**39**	**40**
BeamformIt!, 5ch	-	-	1.350	1.292	1.326	31	36	35
CGMM-EM, 6ch	-	-	1.635	1.483	1.468	48	42	38

7. Conclusion

In this paper, we proposed a resource-efficient linear regression architecture for speech mask estimation as alternative to NNs. Our system uses the dominant eigenvector of the PSD of the microphone signals as feature vector. We compared our results against the most recent model-based and data-driven approaches using the CHiME4 corpus. We have shown that our model yields good results, both in terms of perceptual speech quality and speech mask prediction error, while using two orders of magnitude fewer parameters than comparable NN models. Unlike existing approaches, our system does not require any information about the array geometry or the characteristics of the speech and noise sources. Our assumptions are that the speaker moves slowly and is located in the near-field of the array, while the non-stationary noise is in the far-field.

8. References

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

[2] L. Pfeifenberger and F. Pernkopf, "Blind source extraction based on a direction-dependent a-priori SNR," in *Interspeech*, May 2014.

[3] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, May 2009.

[4] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, Sep. 2004.

[5] 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.

[6] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, Sep. 2011.

[7] ——, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, Jul. 2010.

[8] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline asr in noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 5210–5214, Mar. 2016.

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

[10] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd chime challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 444–451.

[11] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Interspeech*, Sep. 2016.

[12] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Eigenvector-based speech mask estimation for multi-channel speech enhancement," Signal Processing and Speech Communication Laboratory, Technical University of Graz, Austria, Tech. Rep., Dec. 2016.

[13] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

[14] T. Schrank, L. Pfeifenberger, M. Zöhrer, J. Stahl, P. Mowlaee, and F. Pernkopf, "Deep beamforming and data augmentation for robust speech recognition: Results of the 4th chime challenge," in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016)*, 2016.

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

[16] M. Brandstein and D. Ward, *Microphone Arrays*. Berlin–Heidelberg–New York: Springer, 2001.

[17] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process*, pp. 260–275, Feb. 2010.

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

[19] B. D. V. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, no. 5, pp. 4–24, Apr. 1988.

[20] 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.

[21] E. Warsitz, A. Krueger, and R. Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 73–76, May 2008.

[22] L. Pfeifenberger and F. Pernkopf, "A multi-channel postfilter based on the diffuse noise sound field," in *European Association for Signal Processing Conference 2014*, Jun. 2014.

[23] M. G. Shmulik, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, Aug. 2009.

[24] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "DNN-based speech mask estimation for eigenvector beamforming," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017.

[25] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," *Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 430–437, 2012.

[26] "ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2000.