DNN-BASED CONCURRENT SPEAKERS DETECTOR AND ITS APPLICATION TO SPEAKER EXTRACTION WITH LCMV BEAMFORMING

Shlomo E. Chazan, Jacob Goldberger and Sharon Gannot

Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel.
{Shlomi.Chazan,Jacob.Goldberger,Sharon.Gannot}@biu.ac.il

ABSTRACT

In this paper, we present a new control mechanism for LCMV beamforming. Application of the LCMV beamformer to speaker separation tasks requires accurate estimates of its building blocks, e.g. the noise spatial cross-power spectral density (cPSD) matrix and the relative transfer function (RTF) of all sources of interest. An accurate classification of the input frames to various speaker activity patterns can facilitate such an estimation procedure. We propose a DNN-based concurrent speakers detector (CSD) to classify the noisy frames. The CSD, trained in a supervised manner using a DNN, classifies noisy frames into three classes: 1) all speakers are inactive - used for estimating the noise spatial cPSD matrix; 2) a single speaker is active - used for estimating the RTF of the active speaker; and 3) more than one speaker is active - discarded for estimation purposes. Finally, using the estimated blocks, the LCMV beamformer is constructed and applied for extracting the desired speaker from a noisy mixture of speakers.

Index Terms— DNN, multi-speaker detector, LCMV beamformer

1. INTRODUCTION

In recent years we have witnessed an increasing research interest in multi-microphone speech processing due to the rapid technological advances, most notable the introduction of smart assistants for home environment. Adverse acoustic environments are characterized by noise, reverberation and competing speakers. Separating the desired speaker from a mixture of several speakers, while minimizing the noise power, as well as dereverberating the desired source, is therefore a major challenge in the field. A plethora of methods for speaker separation and speech enhancement using microphone arrays can be found in [1, 2].

In this work, we focus on linearly constrained minimum variance (LCMV) beamforming for speaker separation. The LCMV-beamformer (BF) was successfully applied in speech enhancement tasks with multiple signals of interest [3]. The LCMV criterion minimizes the noise power at the BF output while satisfying a set of linear constrains, such that the desired source is maintained while the interfering signals are suppressed. The LCMV-BF can be designed by using the relative transfer functions (RTFs), defined as the ratio of the two acoustic transfer functions (ATFs) relating a source signal and a pair of microphones [4]. The algorithm in [3] demonstrates high separation capabilities and low distortion of the desired source (which are essential for high speech intelligibility and low word error rate of automatic speech recognition systems), provided that the speakers obey a specific activity pattern, namely that time segments when each of the speakers of interest is active alone, can be found. These time segments are utilized for estimating the respective RTFs of all speakers. Time segments with no active speakers are utilized for estimating the noise statistics, another essential component in the beamformer design. An automatic mechanism for determining the activity of the sources of interest is therefore crucial for proper application of the LCMV-BF. An off-line and online estimators of the activities of the speakers were presented in [5] and [6], respectively. In [7] the speaker indexing problem was tackled by first applying a voice activity detector and then estimating the direction of arrival.

Spatial clustering of time-frequency bins and speech presence probability (SPP) estimation techniques were extensively studied in recent year as a mechanism that facilitates beamforming methods in speech enhancement applications. An SPP scheme for constructing a generalized eigenvalue decomposition (GEVD)-based minimum variance distortionless response (MVDR) beamformer with a postfiltering stage was presented in [8], for enhancing a single speaker contaminated by additive noise. An SPP mask is separately extracted from all channels and then averaged to obtain a time-frequency mask used for estimating the noise spatial cross-power spectral density (cPSD) that is further incorporated into an MVDR-BF [9]. An integrated time-frequency masking using deep neural network (DNN) and a probabilistic spatial clustering is proposed in [10] for estimating the steering vector of an MVDR-BF. In [11], a bi-directional LSTM network that robustly estimates soft masks was proposed. The mask is used by a subsequent generalized eigenvalue beamforming that takes into account the acoustic propagation of the sound source. In [12] a speech and noise masks are estimated for constructing an MVDR-BF integrated with an automatic speech recognition system. Recently, we have proposed an LCMV-BF approach for source separation and noise reduction using SPP masks and speaker position identifier [13]. The latter relies on pre-calibrated RTFs which are unavailable in many important scenarios.

In this paper, we present a practical LCMV beamformer with a post-processing stage. For estimating the components of the BF we utilize a single microphone concurrent speakers detector (CSD) and an adaptive dictionary for associating RTF estimates with speakers.

2. PROBLEM FORMULATION

Consider an array with M microphones capturing a mixture of speech sources in a noisy and reverberant enclosure. For simplicity, we will assume that the mixture comprises one desired speaker and one interference speaker. Extension to more speakers is rather straightforward. Each of the speech signals propagates through the acoustic environment before being picked up by the microphone array. In the short-time Fourier transform (STFT) domain, the desired

We thank the Communication and Devices Group, Intel Corporation for their support and collaboration.

and the interfering sources are denoted $s^d(l,k)$ and $s^i(l,k)$, respectively, where l and k, are the time-frame and the frequency-bin indexes. The ATF relating the desired speaker and the m-th microphone is denoted $h_m^d(l,k)$ and the respective ATF of the interfering source is denoted $h_m^i(l,k)$. The ambient stationary background noise at the m-th microphone is $v_m(l,k)$. The received signals can be conveniently formulated in a vector notation:

$$\mathbf{z}(l,k) = \mathbf{h}^d(l,k)s^d(l,k) + \mathbf{h}^i(l,k)s^i(l,k) + \mathbf{v}(l,k) \quad (1)$$

where:

$$\begin{aligned} \mathbf{z}(l,k) &= [z_1(l,k),\ldots,z_M(l,k)]^T \\ \mathbf{v}(l,k) &= [v_1(l,k),\ldots,v_M(l,k)]^T \\ \mathbf{h}^d(l,k) &= [h_1^d(l,k),\ldots,h_M^d(l,k)]^T \\ \mathbf{h}^i(l,k) &= [h_1^i(l,k),\ldots,h_M^i(l,k)]^T. \end{aligned} \quad (2)$$

Equation (1) can be reformulated using normalized signals [3], to circumvent gain ambiguity problems:

$$\mathbf{z}(l,k) = \mathbf{c}^d(l,k)\tilde{s}^d(l,k) + \mathbf{c}^i(l,k)\tilde{s}^i(l,k) + \mathbf{v}(l,k) \quad (3)$$

where

$$\mathbf{c}^d(l,k) = \left[\frac{h_1^d(l,k)}{h_{\text{ref}}^d(l,k)}, \frac{h_2^d(l,k)}{h_{\text{ref}}^d(l,k)}, \ldots, \frac{h_M^d(l,k)}{h_{\text{ref}}^d(l,k)}\right]^T \quad (4)$$

$$\mathbf{c}^i(l,k) = \left[\frac{h_1^i(l,k)}{h_{\text{ref}}^i(l,k)}, \frac{h_2^i(l,k)}{h_{\text{ref}}^i(l,k)}, \ldots, \frac{h_M^i(l,k)}{h_{\text{ref}}^i(l,k)}\right]^T \quad (5)$$

are the desired and interference RTFs, respectively, and 'ref' is the reference microphone. The normalized desired and interference sources are given by $\tilde{s}^d(l,k) = h_{\text{ref}}^d(l,k)s^d(l,k)$ and $\tilde{s}^i(l,k) = h_{\text{ref}}^i(l,k)s^i(l,k)$, respectively.

The goal of the proposed algorithm is to extract the desired source (as received by the reference microphone), namely $\tilde{s}^d(l,k)$, from the received microphone signals, while suppressing the interference source and reducing the noise level. Since the speakers can change roles, we produce two output signals, one for each source.

3. ALGORITHM

We propose to use the LCMV-BF for the task of extracting the desired speech signal. The main contribution is the derivation of a new control mechanism for updating the various blocks of the LCMV-BF. A new DNN-based concurrent speakers detector (CSD), is used to detect the speakers' activity at each time-frame. Noise-only time-frames are used for updating the noise statistics. Frames that are solely dominated by a single speaker are used for RTF estimation. Frames with multiple concurrent speakers active are not used for updating the BF components. We further propose an adaptive dictionary-based method for associating the estimated RTFs with either the desired or the interference sources.

3.1. Linearly Constrained Minimum Variance

Denote the BF weight vector by $\mathbf{w}(l,k) = [w_1(l,k),\ldots,w_M(l,k)]^T$ and the BF output $\hat{\mathbf{s}}^d(l,k)$:

$$\hat{\mathbf{s}}^d(l,k) = \mathbf{w}^H(l,k)\mathbf{z}(l,k). \quad (6)$$

The BF weights are set to satisfy the LCMV criterion with multiple constraints [14]:

$$\begin{aligned} \mathbf{w}(l,k) = \underset{\mathbf{w}}{\text{argmin}} &\left\{\mathbf{w}^H(l,k)\Phi_{vv}(k)\mathbf{w}(l,k)\right\} \\ \text{subject to} \quad &\mathbf{C}^H(l,k)\mathbf{w}(l,k) = \mathbf{g}(l,k) \end{aligned} \quad (7)$$

where $\mathbf{g}(l,k)$ is the desired response, set in our case to $[1,0]^T$,

$$\mathbf{C}(l,k) = \left[\mathbf{c}^d(l,k), \mathbf{c}^i(l,k)\right] \quad (8)$$

is the RTFs-matrix, and $\Phi_{vv}(k)$ is the noise cPSD matrix assumed time-invariant. The well-known solution to (7) is given by,

$$\begin{aligned} \mathbf{w}_{\text{LCMV}}(l,k) = \Phi_{vv}^{-1}(k)\mathbf{C}(l,k)\cdot \\ [\mathbf{C}^H(l,k)\Phi_{vv}^{-1}(k)\mathbf{C}(l,k)]^{-1}\mathbf{g}(l,k). \end{aligned} \quad (9)$$

To calculate (9), an estimate of the RTFs-matrix $\mathbf{C}(l,k)$ and the noise correlation matrix $\Phi_{vv}(k)$ are required. A plethora of methods for estimating the RTFs can be found in the literature. In this work, we use the GEVD-based method described in [3], that necessitates frames dominated by a single active speaker.[1] A method for updating the noise statistics will be presented in the sequel.

3.2. DNN-based concurrent speakers detector (CSD)

The algorithm utilizes a trained DNN-based concurrent speakers detector (CSD). The (single microphone) CSD classifies each frame of the observed signal to one of three classes as follows:

$$\text{CSD}(l) = \begin{cases} \text{Class } \#1 & \text{Noise only} \\ \text{Class } \#2 & \text{Single speaker active} \\ \text{Class } \#3 & \text{Multi speakers active.} \end{cases} \quad (10)$$

To train the DNN-based CSD, a labeled database is required. To construct the database, we generated 1000 single microphone scenarios. For each scenario, the number of clean speech utterances, randomly drawn from the training set of the TIMIT database [15], was set to $n \in \{0,1,2,3\}$. The activity pattern of each utterance was determined by applying the SPP-based voice activity detector (VAD), described in [13], which takes the speech spectral patterns into account. The label of each frame was obtained by adding all VAD values. Depending on the number of speech signals drawn from the database and their associated activity patterns, the labels take values in the range $\{0,1,2,3\}$. Frames with label 0 were classified to the Class #1, frames with label 1 to the Class #2, and frames with labels 2 and 3 were classified to Class #3. Finally, for generating the speech signals, all utterances and a car noise, drawn from the NOISEX-92 [16] database with SNR=15 dB, were added.

The network architecture consists of 2 hidden layers with 1024 rectified linear unit (ReLU) neurons each. The transfer function of the last layer was set as a softmax function and the cross-entropy loss function was used for training the network. The dropout method was utilized in each layer. The batch-normalization method was applied to accelerate the training phase in each layer. Finally, the adaptive moment estimation (ADAM) optimizer was used. The inputs to the network are the log-spectrum vectors of the noisy signals and their associated classes are the outputs.

3.3. Noise adaptation

For initializing the estimation of $\Phi_{vv}(k)$ we assume that the speech utterance starts with a noise-only segment, consisting of a sufficient number of frames that enables accurate matrix inversion. The initial $\Phi_{vv}(k)$ is then obtained by averaging periodograms:

$$\hat{\Phi}_{vv}(k) = \frac{1}{l_v^{\text{stop}} - l_v^{\text{start}}} \sum_{l=l_v^{\text{start}}}^{l_v^{\text{stop}}-1} \mathbf{z}(l,k)\mathbf{z}^H(l,k) \quad (11)$$

[1] An extension for groups of sources exists, but is beyond the scope of this paper that focuses on the separation of two sources.

where l goes over the frames of this segment. Next, frames which were classified to Class #1 are used for updating the noise statistics by a recursive averaging:

$$\Phi_{vv}(l,k) = \alpha \cdot \Phi_{vv}(l-1,k) + (1-\alpha) \cdot \mathbf{z}(l,k)\mathbf{z}^H(l,k) \quad (12)$$

with α the learning rate factor. No noise adaptation is applied in frames which do not belong to Class #1.

3.4. RTF association

As explained above, frames classified to Class #2, which indicates that only a single speaker is active, may be used for RTF estimation. However, as more than one speaker may be active in each utterance (i.e., the desired and interfering speakers), it remains to associate the estimated RTFs with a specific speaker, as explained in the sequel.

We propose a new online RTF association scheme, which is based on a construction of an adaptive dictionary of RTFs of all the active speakers in the scene. The first sequence of frames[2] classified to Class #2 is used to estimate the first RTF of the dictionary. Once a new sequence of frames, classified as Class #2, is detected, and hence a new RTF estimate $\hat{\mathbf{c}}(l,k)$ becomes available, the similarity index (per frequency) between the new RTF estimate and all RTF entries in the dictionary is calculated:

$$S^p(l,k) = \frac{\left|\hat{\mathbf{c}}^H(l,k) \cdot \mathbf{c}^p(k)\right|}{\|\hat{\mathbf{c}}(l,k)\| \cdot \|\mathbf{c}^p(k)\|}. \quad (13)$$

where $p = 1, \ldots, P$ is the entry index and P is the maximum number of different speakers expected in the scene ($P = 2$ in our case). The frequency-wise similarity indexes are then aggregated yielding a frame-wise similarity index:

$$S^p(l) = \sum_{k=0}^{K-1} S^p(l,k) \quad (14)$$

where K is the STFT frame length. The RTF estimate in the l-th frame is finally associated with an existing dictionary entry p_0 or declared as a new entry $p_0 < p_1 \leq P$ according to the following decision rule:

$$\text{RTF association}(l) = \begin{cases} S^{p_0}(l) & \text{if } S^{p_0}(l) > 0.75 \cdot K \\ S^{p_1}(l) & \text{otherwise.} \end{cases} \quad (15)$$

The RTFs dictionary is then updated by either substituting entry p_0 or by adding entry p_1 using the new RTF estimate $\hat{\mathbf{c}}(l,k)$. Note, that if the DNN-based CSD mis-classifies frames with two speakers as a single speaker, the similarity index will be low and therefore no dictionary substitution will occur. An expiration mechanism, together with the upper limit on the number of entries, will guarantee that the wrong estimate will be eventually replaced. In this work, unlike [13], there is no requirement for pre-calibration with known speakers' positions. Here, we present a more flexible scheme that is applicable in many real-life scenarios, e.g. meeting rooms and cars, which is based on two assumptions. First, the speakers in the room are static (slight natural movements allowed). Second, for each speaker in the scene, a sequence of sufficiently long duration for which it is the sole speaker, exists.

Using the estimated RTFs and the noise statistics estimator, as explained above, the LCMV can be constructed. To further improve the interference suppression and the noise reduction, a subsequent postfilter, based on the neural network mixture-maximum (NN-MM) algorithm [17], is applied. The entire algorithm is summarised in Algorithm 1.

[2]To avoid unreliable RTF estimates, only a sequence of 16 consecutive frames is used for the estimation.

Algorithm 1: Summary of the speech enhancement algorithm.

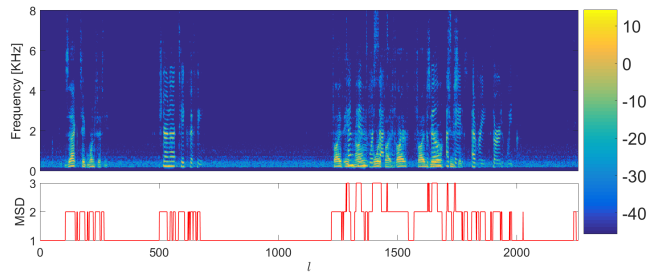
Initialization:
Find Φ_{vv} based on the first 0.5sec. (11)
Input:
Noisy input $\mathbf{z}(l,k)$
for $l = 1 : N_{seg}$ **do**
  Classify frame to one of the three classes (10)
  if *CSD(l)=1* **then**
    Update noise estimation Φ_{vv} (12)
  end
  else if *CSD(l)=2* **then**
    Estimate RTF of the current speaker
    if *First RTF estimation* **then**
      Add to RTF dictionary (8)
    end
    else
      Associate RTF with a speaker (13),(14),(15)
      Update RTF dictionary
    end
  end
  else if *CSD(l)=3* **then**
    continue
  end
end
Enhancement:
  Apply the LCMV-BF $\mathbf{w}_{\text{LCMV}}$ (9) to the noisy input (6)
  Apply NN-MM to the LCMV output [17]

4. EXPERIMENTAL STUDY

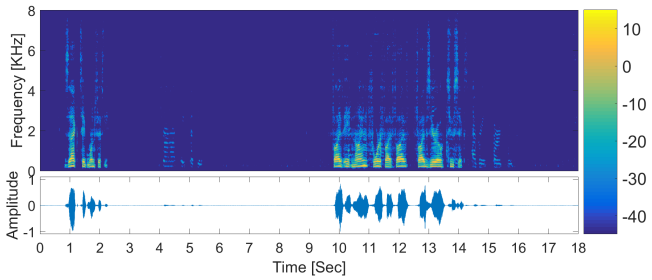
In this section we examine all building blocks of the proposed algorithm and assess the speech quality and intelligibility at the output of the BF. For testing, we have used utterances drawn from the test set of the TIMIT database, as well as recordings of real speakers in a low reverberant environment.

4.1. CSD performance

To test the CSD accuracy, a database was built in a similar way to the training database except that the test set of the TIMIT database was used. Table 1 depicts the confusion matrix of the classifier. It is evident that the CSD correctly detects the noise-only frames with high accuracy (99%). These frames can be used for updating the noise estimation. In addition, when only one speaker is active, the detection rate deteriorates to 75%. The CSD mis-classifies 22% of these frames belonging to Class #3. Although this causes information loss, the consequences are not severe since these frames will be discarded from the estimation procedures of both the noise statistics and the RTFs. Finally, it is also evident that frames with more than one speaker active, are detected by the CSD with 92% accuracy. These frames are not used for any RTF or noise estimation. However, 6% of the multiple speakers active frames were mis-classified as belonging to Class #2. While these frames generates wrong RTF estimates, the measures discussed in Sec. 3.4 may mitigate their harmful consequences.

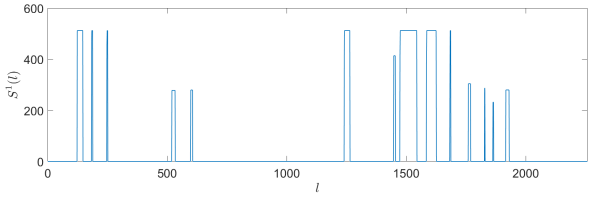


(a) Real scenario with 2 speakers. First speaker #1 is active then speaker #2 and then both.

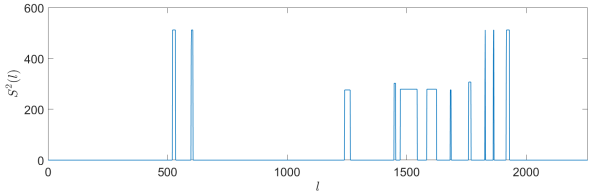


(b) BF output for extracting speaker #1.

Fig. 1: LCMV-BF performance.



(a) Speaker #1



(b) Speaker #2

Fig. 2: RTF association results

Table 1: Confusion matrix of the multiple speaker detector [percentage]

		True		
		1	2	3
	1	99	3	2
Estimated	2	1	75	6
	3	0	22	92

Table 2: Experiment time-line

Time [sec]	0-0.5	0.5-3	3-6	6-9	9-16	16-18
Desired speaker	0	1	0	0	1	0
Interfering speaker	0	0	1	0	1	0
Background noise	1	1	1	1	1	1

4.2. Performance of the LCMV-BF with real recordings

The algorithm performance was also evaluated using a recording campaign carried out in a low reverberant environment.

4.2.1. Setup

The experimental setup is similar to the one in [13]. Speakers can pick there position from four available seats. Microphone array consisting of seven omni-directional microphones arranged in U-shape was used. In order to control the signal to noise ratio (SNR) and the signal to interference ratio (SIR), the signals were separately recorded. Overall, we used 6 speakers (3 male and 3 female speakers) and recorded 1800 utterances. One of the speakers was counting, while the other was reading from the Harvard database [18]. The time-line of signals' activity for all scenarios is described in Table 2.

4.2.2. Sonograms Assessment and CSD classification

Figure 1a depicts an example of the observed signal with SNR=15dB. In the upper panel, the observed signal is depicted and in the lower panel the associated CSD classification results. It can be verified that the noise frames are accurately classified and that Class #2 frames are correctly detected most of the time. The output of the LCMV-BF is depicted in Fig. 1b, clearly indicating the interference suppression capabilities of the proposed algorithm.

4.2.3. RTF association performance

Fig. 2a and Fig. 2b depict the RTF association (14) of each frame to the first and the second speakers, respectively. It is clear that the similarity index $S^1(l)$ is high when the first speaker is active and low when the second speaker is active. The similarity index $S^2(l)$ exhibits opposite trends. Note that both similarity indexes get low values when projected to frames in which both speakers are concurrently active.

4.2.4. STOI results

Figure 3 depicts the short-time objective intelligibility measure (STOI) [19] improvement. We tested different SIR cases from -15dB to 0dB. Each value in the graph is calculated by averaging 18 speech utterances. It is evident that intelligibility of the observed signal is dramatically reduced in low SIR levels, and that the proposed algorithm significantly improves the STOI results.

5. CONCLUSIONS

A new control scheme for LCMV beamforming with two main components was presented: 1) a DNN-based concurrent speakers detector (CSD) for classifying the speech frames into three classes of speakers' activity; and 2) an RTF association procedure based on adaptive dictionary learning. The proposed algorithm was evaluated using signals recorded in natural acoustic environment and exhibits improved results.

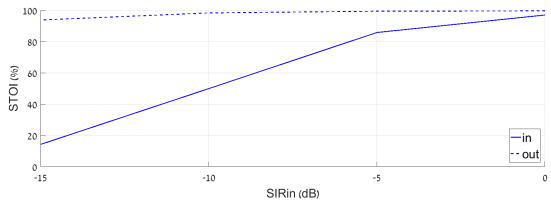


Fig. 3: STOI performance.

6. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1, Springer Science & Business Media, 2008.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[3] S. Markovich, 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, pp. 1071–1086, 2009.

[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, pp. 1614–1626, 2001.

[5] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H/ Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[6] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 499–513, Feb. 2012.

[7] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "Speaker indexing and speech enhancement in real meetings/conversations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 93–96.

[8] 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 (ICASSP)*, 2017, pp. 66–70.

[9] H. Erdogan, J.R. Hershey, S. Watanabe, M.I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Interspeech*, 2016, pp. 1981–1985.

[10] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 286–290.

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

[12] T. Ochiai, S. Watanabe, T. Hori, J.R. Hershey, and X. 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, 2017.

[13] A. Malek, S.E. Chazan, I. Malka, V. Tourbabin, J. Goldberger, E. Tzirkel-Hancock, and S. Gannot, "Speaker extraction using LCMV beamformer with DNN-based SPP and RTF identification scheme," in *The 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, Aug. 2017.

[14] B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE AASP magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[15] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J. G Fiscus, and D.S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27403, 1993.

[16] A. Varga and H.J.M. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, vol. 12, pp. 247–251, Jul. 1993.

[17] S. E. Chazan, J. Goldberger, and S. Gannot, "A hybrid approach for speech enhancement using MoG model and neural network phoneme classifier," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2516–2530, Dec. 2016.

[18] "Harvard database," http://www.cs.columbia.edu/~hgs/audio/harvard.html.

[19] C. H Taal, R. C Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.