LCMV Beamformer with DNN-based Multichannel Concurrent Speakers Detector

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*—Application of the linearly constrained minimum variance (LCMV) beamformer (BF) to speaker extraction tasks in real-life scenarios necessitates a sophisticated control mechanism to facilitate the estimation of the noise spatial cross-power spectral density (cPSD) matrix and the relative transfer function (RTF) of all sources of interest. We propose a deep neural network (DNN)-based multichannel concurrent speakers detector (MCCSD) that utilizes all available microphone signals to detect the activity patterns of all speakers. Time frames classified as no active speaker frames will be utilized to estimate the cPSD, while time frames with a single detected speaker will be utilized for estimating the associated RTF. No estimation will take place during concurrent speaker activity. Experimental results show that the multi-channel approach significantly improves its single-channel counterpart.**

I. Introduction

In recent years we have witnessed an increasing research interest in multi-microphone speech processing due to the rapid technological advances, most notably, the introduction of smart assistants for home environments. 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 at the output, 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], [3].

In this paper we focus on a beamforming method for source extraction, namely the LCMV beamformer (BF) [4], which utilizes RTFs [5] as steering vectors. For a proper application of the LCMV-BF, it is required to blindly determine the activity patterns of the speakers, namely to classify the speech time segments to either no activity, activity of a single source, or concurrent speakers activity. The RTFs of all active speakers can then be estimated during a single active source segments, and the noise statistics can be updated in no active speaker segments.

An offline and online estimators of the activities of the speakers were presented in [6] and [7], respectively. In [8] the speaker indexing problem was tackled by first applying a voice activity detector and then estimating the direction of arrival. In [9], minimum variance distortionless response (MVDR)-BF was used for speech enhancement with a pre-trained dictionary of source location features.

DNN-based signal spectra estimation was used within the expectation-maximization (EM) framework, to estimate a multichannel Wiener filter for separating audio sources [10].

Spatial clustering of time-frequency bins and speech presence probability (SPP) estimation techniques were extensively studied in recent years as a mechanism that facilitates BF methods in speech enhancement applications. An SPP scheme for constructing a generalized eigenvalue decomposition (GEVD)-based MVDR-BF with a postfiltering stage was presented in [11], for enhancing a single speaker contaminated by an 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 cPSD that is further incorporated into an MVDR-BF [12]. An integrated time-frequency masking using DNN and a probabilistic spatial clustering is proposed in [13] for estimating the steering vector of an MVDR-BF. An online MVDR BF based on spatial prior was introduced in [14]. In [15], 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 [16] 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 [17]. The latter relies on pre-calibrated RTFs which are unavailable in many important scenarios.

In [18], a *single microphone* DNN-based concurrent speakers detector (CSD) was presented to control the LCMV-BF. This approach circumvents the pre-calibration requirements. Yet, the spatial information of the microphone array is not taken into account.

In the current paper, we present a *multi-microphone* extension of a DNN-based classifier, called multichannel concurrent speakers detector (MCCSD), for estimating the components of the LCMV-BF.

II. 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.

ISBN 978-90-827970-1-5

In this work, we present a 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.

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 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, respectively. The acoustic transfer function (ATF) relating the desired speaker and the m-th microphone is denoted $h^d_m(l,k)$ and the respective ATF of the interfering source is denoted $h^i_m(l,k)$. 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:

$$\begin{aligned} \mathbf{z}(l,k) &= [z_1(l,k),\ldots,z_M(l,k)]^T \\ \mathbf{v}(l,k) &= [v_1(l,k),\ldots,v_M(l,k)]^T \\ \mathbf{h}^d(l,k) &= [h^d_1(l,k),\ldots,h^d_M(l,k)]^T \\ \mathbf{h}^i(l,k) &= [h^i_1(l,k),\ldots,h^i_M(l,k)]^T. \end{aligned} \quad (2)$$

Equation (1) can be reformulated using normalized signals [4], 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

$$\mathbf{c}^d(l,k) = \left[\frac{h^d_1(l,k)}{h^d_{\text{ref}}(l,k)}, \frac{h^d_2(l,k)}{h^d_{\text{ref}}(l,k)}, \ldots, \frac{h^d_M(l,k)}{h^d_{\text{ref}}(l,k)}\right]^T \quad (4)$$

$$\mathbf{c}^i(l,k) = \left[\frac{h^i_1(l,k)}{h^i_{\text{ref}}(l,k)}, \frac{h^i_2(l,k)}{h^i_{\text{ref}}(l,k)}, \ldots, \frac{h^i_M(l,k)}{h^i_{\text{ref}}(l,k)}\right]^T \quad (5)$$

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

III. Algorithm

As in [18], we use the LCMV-BF for the task of extracting a desired speech signal. A new DNN-based multichannel concurrent speakers detector (MCCSD), is proposed to improve the detection of the speakers' activity at each time-frame. The MCCSD controls the noise statistics update, the RTF estimation and the association of the estimated RTFs with either the desired or the interference sources.

A. DNN-based multichannel concurrent speakers detector (MCCSD)

In [18], a *single microphone* DNN-based CSD was introduced. Each frame of the observed signal was classified to one of three classes as follows:

$$\text{CSD}(l) = \begin{cases} \text{Class } \#1 & \text{Noise only} \\ \text{Class } \#2 & \text{Single speaker active} \\ \text{Class } \#3 & \text{Multi speakers active.} \end{cases} \quad (6)$$

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.

The single microphone DNN-based CSD was trained with a generated labeled database. The database was constructed using the TIMIT database [19]. This approach has some major drawbacks. First, in the TIMIT database the utterances were recorded close to the microphone. In real-life scenarios the speakers are not always close to the microphones. Therefore, the room impulse responses (RIRs) are not taken into account in the training phase of this approach. Additionally, as only a single microphone was used, the spatial information of the microphone array is not utilized.

In this paper we present an improved classification mechanism, namely a DNN-based multichannel concurrent speakers detector (MCCSD). First, to train the multichannel concurrent speakers detector (MCCSD) the recorded database described in [17] was used. This time, the acoustics of the scene is part of the training phase. Two hundred different scenarios were used as the training data. To represent various real-life scenarios, we trained the network with two signal to interference ratio (SIR) levels at $\{0, 5\}$ dB and three signal to noise ratio (SNR) levels $\{5, 10, 15\}$ dB. In addition, unlike [18], in which the log-spectrum of a single microphone was used as the input to the DNN-based CSD, here we concatenate the log-spectrum of all microphones to a longer feature vector.

The network architecture consists of two 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.

B. Linearly Constrained Minimum Variance

The well-known LCMV-BF [20], $\mathbf{w}_{\text{LCMV}}$ is given by,

$$\mathbf{w}_{\text{LCMV}}(l,k) = \Phi^{-1}_{vv}(k)\mathbf{C}(l,k)\cdot [\mathbf{C}^H(l,k)\Phi^{-1}_{vv}(k)\mathbf{C}(l,k)]^{-1}\mathbf{g}(l,k). \quad (7)$$

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

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

ISBN 978-90-827970-1-5 © EURASIP 2018

is the RTFs-matrix, and $\Phi_{vv}(k)$ is the noise cPSD matrix assumed to be time-invariant. The BF is then applied to the noisy input to extract the desired speaker:

$$\hat{\mathbf{s}}^d(l,k) = \mathbf{w}^H_{\text{LCMV}}(l,k)\mathbf{z}(l,k). \quad (9)$$

To calculate (7), an estimate of the noise correlation matrix $\Phi_{vv}(k)$ and the RTFs-matrix $\mathbf{C}(l,k)$ are required. In the following, we describe how the proposed MCCSD is utilized to gain these estimations.

C. Noise adaptation

We first initialize the estimation of $\Phi_{vv}(k)$ with the identity matrix, namely $\mathbf{I}_{M\times M}$. Next, frames which were classified to Class #1 are used for updating the noise statistics by a recursive averaging:

$$\Phi_{vv}(l,k) = \alpha \cdot \Phi_{vv}(l-1,k) + (1-\alpha)\cdot \mathbf{z}(l,k)\mathbf{z}^H(l,k) \quad (10)$$

with α the learning rate factor. The noise adaptation is not applied in frames which do not belong to Class #1.

D. RTF association

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 [4], that necessitates frames dominated by a single active speaker. Consequently, frames classified to Class #2, which indicates that only a single speaker is active, are used.

An RTF dictionary of all the active speakers in the scene is constructed adaptively. The RTF of the first active speaker is estimated with the first sequence of frames classified as Class #2 frames.

In the subsequent Class #2 frames, a new RTF estimate $\hat{\mathbf{c}}(l,k)$ becomes available. To identify the newly acquired RTF with already known RTF, a similarity index (per frequency) between the acquired RTF estimate and all already available RTF entries in the dictionary is calculated:

$$S^p(l,k) = \frac{\left|\hat{\mathbf{c}}^H(l,k)\cdot \mathbf{c}^p(k)\right|}{\|\hat{\mathbf{c}}(l,k)\|\cdot\|\mathbf{c}^p(k)\|} \quad (11)$$

where p is an entry index to the dictionary. The frequency-wise similarity indexes are then aggregated yielding a frame-wise similarity index:

$$S^p(l) = \sum_{k=0}^{K-1} S^p(l,k) \quad (12)$$

where K is the STFT frame length. The RTF estimate in the l-th frame is then either associated with the existing dictionary entry, $p=1$, or declared as a new entry, namely $p=2$. The RTFs dictionary is then updated by either substituting entry $p=1$ with a more accurate RTF estimate, or by adding a new entry $p=2$ using the new RTF estimate $\hat{\mathbf{c}}(l,k)$. This procedure is repeated until the maximum number of expected speakers P has been acquired ($P=2$ in our case).

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 [21], is applied. The entire algorithm is summarized in Algorithm 1.

Algorithm 1: LCMV-BF with the MCCSD.

Input:
Noisy input in the STFT domain $\mathbf{z} = \{\mathbf{z}(l,k)\}$
for $l = 1 : N_{seg}$ **do**
  Classify frame l to one of the three classes (6).
  if *MCCSD(l)=1* **then**
    Update noise estimation Φ_{vv} (10).
  end
  else if *MCCSD(l)=2* **then**
    Update the RTF dictionary using (11),(12).
  end
  else if *MCCSD(l)=3* **then**
    continue
  end
end
Enhancement:
  Apply the LCMV-BF $\mathbf{w}_{\text{LCMV}}$ (7)
  to the noisy input (9).
  Apply the NN-MM algorithm [21] to the LCMV output.

IV. Experimental study

In this section we compare the performance of the new MCCSD approach and the single-channel CSD [18].

A. Setup

The experimental setup is similar to the one in [17]. Speakers can pick their position from four available seats. A microphone array consisting of seven omni-directional microphones arranged in U-shape was used. In order to control the SNR and the 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 [22]. The time-line of signals' activity for all scenarios is described in Table I. Both approaches have the same network architecture (except the input size), and are trained on the same database. For training data, we used 200 different scenarios of the database. The CSD was trained on the reference microphone only, and the MCCSD was trained with all the microphones as described in Sec. III-A. For testing, 80 scenarios, different from the scenarios used for training, were used.

B. Classification performance

To test the advantage of the new MCCSD approach over the single-channel CSD approach, we first compared the accuracy

TABLE I: 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

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE II: Confusion matrix of CSD [percentage].

Estimated \True	1	2	3
1	94.9	13.4	0
2	5.1	76.9	38.9
3	0	9.7	61.1

TABLE III: Confusion matrix of MCCSD [percentage].

Estimated \True	1	2	3
1	98.1	7.9	0
2	1.9	83.7	16.8
3	0	8.4	83.2

of their classification. Note that unlike [18], here we test the classification performance on real speakers recordings and not on the TIMIT database. The utterances are tested in two SIR levels $\in \{0, 5\}$ dB and in three SNR levels $\in \{5, 10, 15\}$ dB.

Table II and Table III depict the confusion matrices of the CSD and MCCSD classifiers, respectively. It is clear that both approaches correctly detect the noise-only frames with high accuracies (MCCSD slightly better than the CSD). As mentioned, these frames are used for updating the noise statistics estimation. When only one speaker is active, the CSD detection rate deteriorates to 76.9% while mis-classifying 9.7% of these frames as belonging to Class #3 and 13.4% as belonging to class #1. The proposed MCCSD outperforms the CSD and improves the detection rate to 83.7.

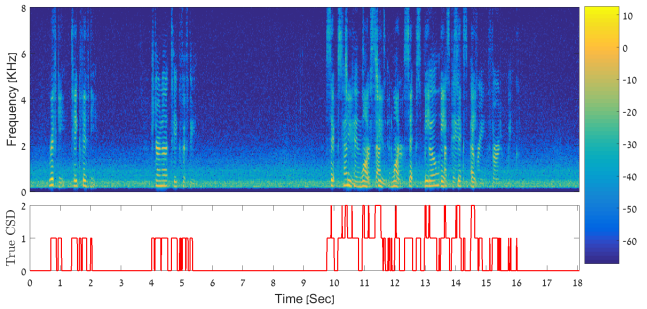
Finally, frames with more than one active speaker, are detected by the CSD with 61.1% accuracy. These frames are not used for any RTF or noise estimation. However, 38.9% of the multiple speakers active frames were mis-classified as belonging to Class #2. These frames may generates wrong RTF estimates. The proposed MCCSD approach significantly reduces the mis-classification to only 16.8%.

Clearly, the MCCSD classifier significantly outperforms the single-channel CSD. The dramatic improvement is a result of the utilization of the spatial information. Consequently, the LCMV-BF parameters are more accurate, and the enhancement is expected to improve.

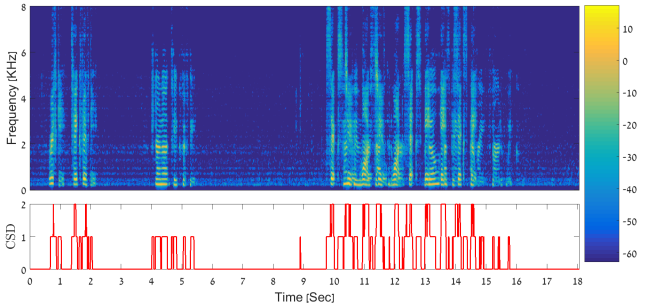
C. Performance of the LCMV-BF with the CSD and the MCCSD

We next compare the contribution of the two classifiers to the overall performance of the BF.

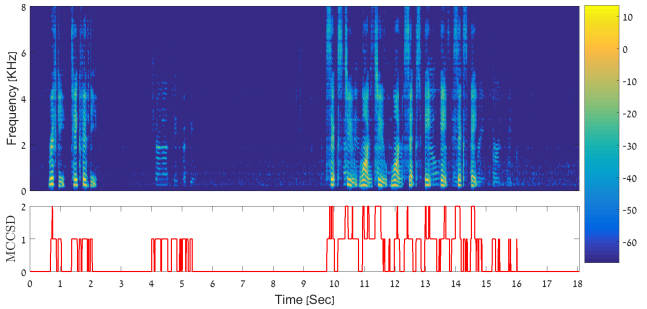
1) Sonograms Assessment: Figure 1a depicts an example of the observed noisy signal with SNR=5dB and SIR=0dB. In the upper panel, the observed signal is depicted and in the lower panel the associated true CSD. The output of the LCMV-BF with the CSD classifier is depicted in Fig. 1b in the upper panel, and the CSD classification results in the lower panel. Similarly, Fig. 1c depicts the output of the LCMV-BF and the MCCSD classifier results. It is evident that the MCCSD classification performance is more accurate than of the CSD. Consequently, the extraction of the desired speaker as well as the noise reduction are significantly better with the MCCSD.



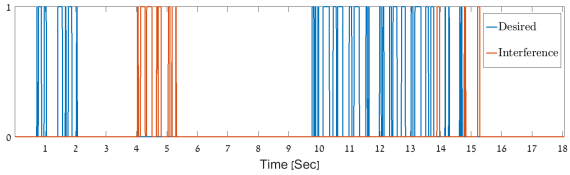
(a) Real scenario with 2 speakers. First speaker #1 is active then speaker #2 and then both.



(b) BF output for extracting speaker #1 with the CSD.



(c) BF output for extracting speaker #1 with the MCCSD.



(d) RTF association of the MCCSD.

Fig. 1: LCMV-BF performance with the two approaches.

2) RTF association performance: We continue with the same scene as in the former section. Fig. 1d depicts the algorithm decisions regarding the pertinence of each frame to one of the speakers in the RTF matrix (8). The desired speaker is designated with a blue line, and the interference speaker with a red line, where '1' denotes speaker active and '0' for speaker inactive. It is clear that the algorithm decisions are very accurate. The frames are well associated with the real active speakers. Note, that within the frames of concurrent speakers activity, the algorithm accurately finds frames dominated by one of the signals.

ISBN 978-90-827970-1-5 © EURASIP 2018

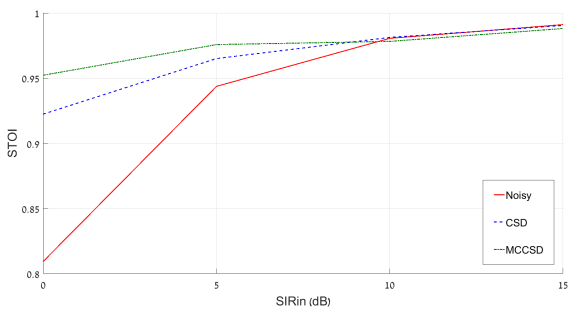


Fig. 2: STOI performance.

3) STOI results: Figure 2 depicts the comparison between the short-time objective intelligibility measure (STOI) [23] results at the LCMV-BF output while using either the CSD or the proposed MCCSD. We tested different SIR cases from 0dB to 15dB. Each value in the graph is calculated by averaging 20 speech utterances. The intelligibility of the observed signal at SIR $= 0$ dB, for example drops to approximately 80%. While the LCMV-BF with the CSD improves the STOI performance to approximately 92%. Using MCCSD further improves the STOI results to approximately 96%.

V. Conclusions

A new multichannel control scheme for LCMV beamforming with two main components was presented: 1) a DNN-based multichannel concurrent speakers detector (MCCSD) 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.

References

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*, vol. 615, Springer, 2007.

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

[3] 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.

[4] 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.

[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, pp. 1614–1626, 2001.

[6] 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.

[7] 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.

[8] 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.

[9] S. Araki, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, T. Higuchi, T. Yoshioka, D. Tran, S. Karita, and T. Nakatani, "Online meeting recognition in noisy environments with time-frequency mask based MVDR beamforming," in *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*. IEEE, 2017, pp. 16–20.

[10] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, Sept. 2016.

[11] 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.

[12] 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.

[13] 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.

[14] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[15] 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.

[16] 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.

[17] 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.

[18] S. E. Chazan, J. Goldberger, and S. Gannot, "DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming," in *IEEE International Conference on Audio and Acoustic Signal Processing (ICASSP)*, 2018.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

ISBN 978-90-827970-1-5 © EURASIP 2018