# Speaker Extraction using LCMV Beamformer with DNN-based SPP and RTF Identification Scheme

Ariel Malek[1], Shlomo E. Chazan[1], Ilan Malka[2], Vladimir Tourbabin[2], Jacob Goldberger[1], Eli Tzirkel-Hancock[2], and Sharon Gannot[1]
[1]Faculty of Engineering, Bar-Ilang University, Ramat-Gan, 5290002, Israel
[2]General Motors Advanced Technical Center Israel
Corresponding author: Sharon.Gannot@biu.ac.il

***Abstract*—The linearly constrained minimum variance (LCMV)-beamformer (BF) is a viable solution for desired source extraction from a mixture of speakers in a noisy environment. The performance in terms of speech distortion, interference cancellation and noise reduction depends on the estimation of a set of parameters. This paper presents a new mechanism to update the parameters of the LCMV-BF. A new speech presence probability (SPP)-based voice activity detector (VAD) controls the noise covariance matrix update, and a speaker position identifier (SPI) procedure controls the relative transfer functions (RTFs) update. A postfilter is then applied to the BF output to further attenuate the residual noise signal. A series of experiments using real-life recordings confirm the speech enhancement capabilities of the proposed algorithm.**

## I. Introduction

Enhancing noisy signals, even if recorded in low echoic enclosure, is a cumbersome task, due to the co-existence of interfering speakers and background noise contaminating the desired speaker. In this work, we focus on a multi-microphone solution for extracting the desired speaker from a noisy mixture of multiple speakers.

Beamforming is a widely-used method for speech enhancement using microphone arrays [1]. A comprehensive overview of various speech enhancement and source extraction algorithms is presented in [2]. The minimum variance distortionless response (MVDR)-BF [3], [4] steers a beam towards the desired source such that the desired signal remains undistorted while minimizing the other noise signals. Yet, if interfering speaker is active and a background noise is also present, the MVDR-BF might not effectively mitigate an interfering speaker.

The LCMV-BF was successfully applied in speech enhancement tasks with multiple signals of interest [5]. 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 blocked. The LCMV-BF and the MVDR-BF can be designed by using the RTFs[1] [4], [5], rather than a simple steering vector, which is based on the time difference of arrival (TDOA) between microphone pairs, to guarantee sufficiently high performance measures in a wide range of reverberation levels.

[1]The RTF is defined as the ratio of the two acoustic transfer functions (ATFs) relating a source signal and a pair of microphones

There are many scenarios for which both the desired speaker and the interfering speakers are located in approximately fixed positions, e.g. around a table in a conference room. Nordholm et al. [6], [7] proposed to use either the MVDR-BF, implemented in a general sidelobe canceller (GSC) structure, or the multichannel Wiener filter to extract a single desired source. Both solutions are only capable of suppressing the background noise and are not aiming at the cancellation of a competing speaker. Calibration stage may be beneficial if the speakers' positions are approximately constant. In [8], the authors used recorded signals in the calibration stage in order to find the BF parameters, and in the test phase they used a "master-slave" structure in which the BF weights of the "master" BF, applied to the real signals, are copied from the "slave" BF, learned from the pre-recorded signals.

The LCMV-BF is a more suitable solution in multiple concurrent speakers scenarios. To apply the LCMV-BF, it is necessary to estimate the RTFs and the noise covariance matrix. In this paper, we present a practical end-to-end implementation of a multichannel speech enhancement system based on the LCMV beamformer and a post-processing stage. A recently proposed neural network mixture-maximum (NN-MM) algorithm [9] is utilized to derive a voice activity detector (VAD). A new speaker position identifier (SPI) is proposed based on a pre-trained RTF library. Finally, the NN-MM algorithm is applied to the LCMV output as a postfilter.

## II. Method

### A. Problem Formulation

Consider an array with $M$ microphones in a predefined fixed position. The array captures a desired speaker contaminated by interfering speaker and stationary background noise. Each of the involved signals undergo filtering by the acoustic impulse response (AIR) before being picked up by the microphones. In the time-domain the signal received by the $m$-th microphone is given by:

$$z_m(n) = s^d(n) * h^d_m(n) + s^i(n) * h^i_m(n) + v_m(n) \qquad (1)$$

where $s^d(n)$, $s^i(n)$ and $v_m(n)$ are the desired source, the interfering source and the stationary background noise, respectively. In real-life scenarios more than two speakers can be simultaneously active. In this paper, for simplicity, we focus on the two speakers scenario. The AIR between the desired

ISBN 978-0-9928626-7-1 © EURASIP 2017

speaker and the $m$-th microphone is $h_m^d(n)$ and similarly, the AIR between the interfering source and the $m$-th microphone is $h_m^i(n)$. In the short-time Fourier transform (STFT) domain $z_m(n)$ can be stated as:

$$z_m(l,k) = s^d(l,k)\cdot h_m^d(l,k) + s^i(l,k)\cdot h_m^i(l,k) + v_m(l,k) \quad (2)$$

where $l$ and $k$ are the time-frame and the frequency indexes, respectively. The terms $h_m^d(l,k)$ and $h_m^i(l,k)$ are the ATFs, defined as the Fourier transform of the corresponding AIRs. The received signals in (2) can be conveniently formulated in a vector notation:

$$\begin{aligned}\mathbf{z}(l,k) =& \mathbf{h}^d(l,k)s^d(l,k) + \mathbf{h}^i(l,k)s^i(l,k) + \mathbf{v}_m(l,k) \\ =& \mathbf{H}(l,k)\mathbf{s}(l,k) + \mathbf{v}(l,k)\end{aligned} \quad (3)$$

where:

$$\begin{aligned}\mathbf{z}(l,k) &= [z_1(l,k),\ldots,z_M(l,k)]^T \\ \mathbf{v}(l,k) &= [v_1(l,k),\ldots,v_M(l,k)]^T \\ \mathbf{h}^d(l,k) &= [h_1^d(l,k),\ldots,{}_M^d(l,k)]^T \\ \mathbf{h}^i(l,k) &= [h_1^i(l,k),\ldots,{}_M^i(l,k)]^T \\ \mathbf{H}(l,k) &= [\mathbf{h}^d(l,k),\mathbf{h}^i(l,k)] \\ s(l,k) &= [s^d(l,k),\mathbf{s}^i(l,k)]^T.\end{aligned} \quad (4)$$

Assuming the desired speech signals, the interference and the noise signals are uncorrelated, the correlation matrix of the received signals is given by:

$$\Phi_{zz}(l,k) = \Phi_{dd}(l,k) + \Phi_{ii}(l,k) + \Phi_{vv}(l,k) \quad (5)$$

with:

$$\begin{aligned}\Phi_{dd}(l,k) &= \left[\sigma^d(l,k)\right]^2\mathbf{h}^d(l,k)\left[\mathbf{h}^d(l,k)\right]^H \\ \Phi_{ii}(l,k) &= \left[\sigma^i(l,k)\right]^2\mathbf{h}^i(l,k)\left[\mathbf{h}^i(l,k)\right]^H \\ \Phi_{vv}(k) &= {\sigma_v}^2 I_{M\times M}\end{aligned} \quad (6)$$

and $(\cdot)^H$ is the conjugate-transpose operation.

### B. System overview

In this paper, we propose a new control mechanism for the update of the LCMV-BF parameters. The noise covariance matrix is initialized by averaging the first frames of the utterance, assumed to be noise-only frames. The NN-MM algorithm [9] is then applied to the reference microphone to extract an SPP map. A VAD is calculated based on the SPP detector and used to control the noise estimation update. In the calibration stage, an RTF library, consisting of a specific RTF for each position, is calculated. In the test stage, the RTFs-matrix is initialized with the RTFs of this library. A new scheme for SPI is also proposed to classify speech-active frames to be either associated with the desired or interfering sources. The classification results of the SPI control the RTFs-matrix update. Finally, the LCMV is applied to the noisy input, followed by a postfilter based on the NN-MM algorithm [9].

### C. Linearly Constrained Minimum Variance

In this work, we are interested in extracting the desired speaker from the noisy signal, while suppressing the interference signal. For that, we are applying a BF $\mathbf{w}(l,k)$ to the noisy signal $\mathbf{z}(l,k)$. The BF output $\hat{\mathbf{s}}^d(l,k)$ is given by:

$$\hat{\mathbf{s}}^d(l,k) = \mathbf{w}^H(l,k)\mathbf{z}(l,k) \quad (7)$$

where $\mathbf{w}(l,k) = [w_1(l,k),\ldots,w_M(l,k)]^T$.

The filters are set to satisfy the LCMV criterion with multiple constraints [10]:

$$\begin{aligned}\mathbf{w}(l,k) = \underset{\mathbf{w}}{\operatorname{argmin}} &\left\{\mathbf{w}^H(l,k)\Phi_{vv}\mathbf{w}(l.k)\right\} \\ \text{subject to} \quad &\mathbf{C}^H(l,k)\mathbf{w}(l,k) = \mathbf{g}(l,k)\end{aligned} \quad (8)$$

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

$$\mathbf{C}(l,k) = [\mathbf{c}_d(l,k),\mathbf{c}_i(l,k)] \quad (9)$$

is the RTFs-matrix, with

$$\mathbf{c}_d(l,k) = \left[\frac{h_1^d(l,k)}{h_{\text{ref}}^d(l,k)},\frac{h_2^d(l,k)}{h_{\text{ref}}^d(l,k)},\ldots,\frac{h_M^d(l,k)}{h_{\text{ref}}^d(l,k)}\right]^T \quad (10)$$

and

$$\mathbf{c}_i(l,k) = \left[\frac{h_1^i(l,k)}{h_{\text{ref}}^i(l,k)},\frac{h_2^i(l,k)}{h_{\text{ref}}^i(l,k)},\ldots,\frac{h_M^i(l,k)}{h_{\text{ref}}^i(l,k)}\right]^T \quad (11)$$

where 'ref' is the reference microphone. The well-known solution to (8) is given by,

$$\begin{aligned}\mathbf{w}_{\text{LCMV}}(l,k) = &\Phi_{vv}^{-1}(l,k)\mathbf{C}(l,k)\times \\ &[\mathbf{C}^H(l,k)\Phi_{vv}^{-1}(l,k)\mathbf{C}(l,k)]^{-1}\mathbf{g}(l,k).\end{aligned} \quad (12)$$

To calculate (12), an estimate of the RTFs-matrix $\mathbf{C}(l,k)$ and the noise correlation matrix $\Phi_{vv}$ are required.

### D. Noise estimation

In order to estimate $\Phi_{vv}$, we assume that there are time segments for which none of the speakers is active. These segments are utilized for estimating the stationary noise power spectral density (PSD). Define $\left[l_v^{\text{start}},l_v^{\text{stop}}\right]$ as a noise-only time segment and initialize the corresponding PSD matrix:

$$\hat{\Phi}_{vv}(k) = \frac{1}{l_v^{\text{stop}} - l_v^{\text{start}}}\sum_{l=l_v^{\text{start}}}^{l_v^{\text{stop}}-1}\mathbf{z}(l,k)\mathbf{z}^H(l,k). \quad (13)$$

We assume that the first $0.5$ sec can be utilized for initializing $\Phi_{vv}$, and discuss a VAD-based adaptation scheme in Sec. III.

### E. RTF estimation

This section is dedicated to the estimation of the RTFs-matrix $\mathbf{C}(l,k)$ (9). In static scenarios, the RTFs of the desired and the interfering sources can be pre-estimated in a calibration stage. Under the assumption that the sources' positions are approximately fixed, their identity as either desired or interference source can be determined. The RTFs library can be constructed off-line using different speakers and utterances

ISBN 978-0-9928626-7-1 © EURASIP 2017

than used in the test stage. As the match between the pre-estimated RTFs and the corresponding RTFs of the actual environment is presumably high, good identification results may be obtained.

It is assumed that there are time-frames in which only one source (either desired or interference) is active. These frames, $[l^{\text{start}}, l^{\text{stop}}]$, can be identified and classified to desired/interference segment as described in Sec. III. This segment can then be used for estimating the corresponding RTF. This assumption, although restrictive, can be met in realistic scenarios, for which double-talk scenarios only rarely occurs. Now, applying the generalized eigenvalue decomposition (GEVD) to $\hat{\Phi}_{zz}(l,k)$ and the stationary-noise PSD matrix $\hat{\Phi}_{vv}(l,k)$ we have:

$$\hat{\Phi}_{zz}(l,k) = \lambda(k)\hat{\Phi}_{vv}(l,k)\mathbf{f}(k) \tag{14}$$

where $\lambda(k)$ is the generalized-eigenvalue, $\mathbf{f}(k)$ is the corresponding generalized-eigenvector and $\hat{\Phi}_{zz}(l,k)$ is the correlation matrix estimated from the frames $[l^{\text{start}}, l^{\text{stop}}]$ and (13). Under the assumption that only one speaker is active at this segment, the eigenvector associated with the largest eigenvalue is a scaled version of an RTF $\mathbf{c}(l,k) \in \{\mathbf{c}_d(l,k), \mathbf{c}_i(l,k)\}$. Therefore, we normalize the result to obtain a proper RTF:

$$\hat{\mathbf{c}}(l,k) = \frac{\hat{\Phi}_{vv}(l,k)\mathbf{f}(k)}{\left[\hat{\Phi}_{vv}(l,k)\mathbf{f}(k)\right]_{\text{ref}}}. \tag{15}$$

### F. DNN-based SPP post filter

The output of the BF $\hat{s}^d(l,k)$ (7) is a single channel signal contaminated by residual noise. We apply the NN-MM algorithm [9] to the BF output to enhance the noisy signal. The NN-MM algorithm utilizes a phoneme-based Mixture of Gaussians (MoG) where each Gaussian represents a different phoneme, and a trained deep neural network (DNN) phoneme-classifier, which classifies time-frames to one of the phonemes in the phoneme-based MoG. By merging the generative MoG and the discriminative DNN, the NN-MM constructs a time-frequency SPP map $\rho(l,k)$. A soft spectral attenuation, which was found useful for speech enhancement [9], [11], is then applied to the BF output:

$$\tilde{s}^d(l,k) = \hat{s}^d(l,k) - (1-\rho(l,k)) \cdot \beta \tag{16}$$

where $\beta$ is the soft attenuation level. Note, that (16) is carried out in the log-spectrum domain.

## III. The LCMV control mechanisms

Until now it was assumed that the time-segments in which each speaker is active and their classification as desired/interference source are known. In this section we describe control mechanisms that will be utilized to infer this information from the measured data in real-life scenarios.

### A. SPP-based VAD

As was mentioned above, the noise PSD matrix $\Phi_{vv}(l,k)$ is a crucial component in the LCMV BF design. In (13) the time-frames in which only the background noise is active are required. Here, we propose an SPP-based VAD to determine these noise-only frames. The noisy signal from the reference microphone, denoted $z_{\text{ref}}(l,k)$, is used as the input to the NN-MM algorithm. The NN-MM calculates the SPP of the noisy signal, $\rho(l,k)$. The probabilities are then aggregated across frequencies to yield a VAD decision per frame:

$$V(l) = \begin{cases} 1 & \sum_k \rho(l,k) > T_r \\ 0 & \sum_k \rho(l,k) \leq T_r \end{cases} \tag{17}$$

where $T_r$ is the threshold value. In our implementation we set $T_r = N_{\text{DFT}}/4$, where $N_{\text{DFT}}$ is the STFT frame-length.

Note that the proposed VAD is set to the value '1' if any speech source is active, regardless of the identity of this source. Given that the current frame is noise-dominant, the noise estimation can be recursively updated:

$$\Phi_{vv}(l,k) = \alpha \cdot \Phi_{vv}(l-1,k) + (1-\alpha) \cdot \mathbf{z}(l,k)\mathbf{z}^H(l,k) \tag{18}$$

where $\alpha$ is the learning rate factor. Otherwise, no noise adaptation is applied.

### B. Speaker position identification based on pre-trained RTFs

An accurate RTF estimation is a crucial component in the BF design. For that, time-frames dominated by a single speaker should be determined. Given the VAD in (17), we know whether any speech source is active or not. Yet, we do not know whether the desired speaker, the interference speaker or both are active.

In our scenario, the speakers position are fixed, and the reverberation level is low. The fixed positions makes the pre-training stage feasible. Consequently, in the calibration stage an RTF library, which consists of a specific RTF for each position, is measured. We set the components of the RTF library to $\mathbf{c}^s(k)$, $s = 1, \ldots, N_s$, where $N_s$ is the number of possible positions (in our experiments $N_s = 4$). This stage is only carried out once and does not recur. The RTFs-matrix (9), is then initialized with the components from the library, without loss of generality it is set to $\mathbf{C}(k) = \left[\mathbf{c}^1(k), \mathbf{c}^4(k)\right]$. At the test phase, given speech-active frames, an RTF is estimated using (15). The estimated RTF $\hat{\mathbf{c}}(l,k)$ is then projected to each component of the RTF library, by calculating the *cosine distance*:

$$D_i(l,k) = \frac{\left|\hat{\mathbf{c}}(l,k)^H \cdot \mathbf{c}^i(k)\right|}{\|\hat{\mathbf{c}}(l,k)\| \cdot \|\mathbf{c}^i(k)\|}. \tag{19}$$

The fact that we deal with low reverberation level makes the distance measure (19) a valid affinity measure between impulse responses (see discussion in [12].). Under the assumption that only one speaker is active, the position of the active speaker is determined by $I(l)$:

$$I(l) = \underset{i}{\text{argmax}} \sum_k D_i(l,k). \tag{20}$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

Given the speaker position, the RTFs-matrix (9) is updated with the current RTF. Note, that if both speakers are active, the cosine distance will be smaller than 1 for each position. Consequently, these time-frames will not be utilized for updating the RTFs-matrix. The algorithm is summarized in Alg. 1.

**Algorithm 1:** Speech enhancement algorithm.

**Initialization:**
Find $\Phi_{vv}$ based on the first $0.5$sec. (13).
Set $\mathbf{C}(l,k) = \left[\mathbf{c}^1(l,k), \mathbf{c}^4(l,k)\right]$ (9).
**Input:**
Noisy input $\mathbf{z}(l,k)$.
**for** $l = 1 : N_{seg}$ **do**
  Calculate SPP-based VAD utilizing NN-MM (17).
  **if** *Noise* **then**
    Update noise estimation $\Phi_{vv}$ (18).
  **end**
  **else if** *Speaker active* **then**
    1) Estimate RTF of the current speaker (15).
    2) Calculate cosine distance (19).
    3) Determine which speaker is active (20):
    **if** *Desired speaker* **then**
      Update $\hat{\mathbf{c}}^d(l,k)$ in (9).
    **end**
    **else if** *Interfering speaker* **then**
      Update $\hat{\mathbf{c}}^i(l,k)$ in (9).
    **end**
    **else if** *Two speakers active* **then**
      continue.
    **end**
  **end**
  **Enhancement:**
    Find $\mathbf{w}_{\text{LCMV}}$ (12).
    Apply beamforming on the noisy input (7).
    Apply NN-MM to the LCMV output (16).
**end**

## IV. Experimental study

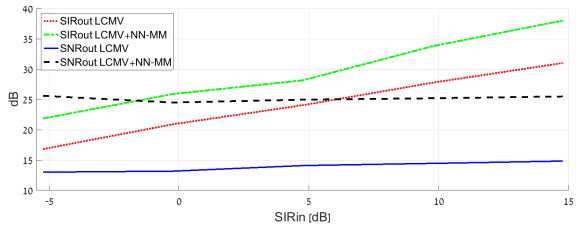
### A. Experiment setup

The algorithm performance was evaluated by a set of experiments using a recording campaign carried out in a low echoic enclosure. There are four positions in the experiment, denoted $1, 2, 3, 4$. Microphone array consisting of seven omni-directional microphones arranged in U-shape was used.

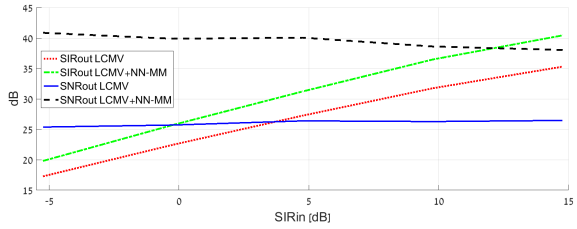
In order to control the signal to noise ratio (SNR) and the signal to interfering ratio (SIR), the desired speaker, the interfering speakers and the background noise were separately recorded. The desired speaker was located at position #1. The fifth microphone was chosen to be the reference microphone. The other positions were occupied with the interfering speaker. The background noise was recorded separately. For the recording campaign, we used 6 speakers (3 male and 3 female speakers) and recorded 1800 utterances. The desired speaker was counting, while the interfering speakers were reading from the Harvard database [13]. The separate recordings were then used to synthesize real-life scenarios. The time-line for each scenario is described in Table I, and explained in the sequel. At $0 \div 0.5$sec the speakers are inactive, at $0.5 \div 3$sec the desired speaker speaks alone, at $3 \div 6$sec only the interference speaker is active, at $6 \div 9$sec the sources are not active and at $9 \div 16$sec the desired speaker and the interfering speaker are all active. Note, that the background noise is present during the entire utterance.

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 |




(a) SNRin = -5dB.




(b) SNRin = 10dB.

Fig. 1: SNRout and SIRout as a function of SIRin.

### B. Experimental results

To evaluate the enhancement capabilities, we evaluated the SNRout and the SIRout at the output of the algorithm as a function of SIRin at the input to the algorithm in the range $\{-5, 0, 5, 10, 15\}$dB for SNRin in the range $\{-5, 0, 5, 10\}$dB. In Fig. 1 we present the results obtained by averaging of 15 signals for each scenario. Due to space constrains we only present the results for SNRin=-5,10dB. The resulting SNRout and SIRout are presented with and without the postfilter. It is easily verified that the SIRout is approximately linearly growing with SIRin. Additionally, the NN-MM postfilter significantly improves the SNRout. This is consistent for all levels of SNRin. Interestingly, the postfilter also improves SIRout. This may be attributed to the spectral shape distortion introduced to the interference source resulting by the application of the LCMV-BF. To further evaluate the performance of the algorithm we set SIRin to 5dB, and the SNRin to 2dB. The desired speaker was placed in the position #1, and the interfering speaker was placed in position #3. The noisy STFT at the reference microphone $z_{\text{ref}}$ is presented in Fig. 2. The decisions of the SPP-based VAD are marked with a red line on top of the time

ISBN 978-0-9928626-7-1 © EURASIP 2017

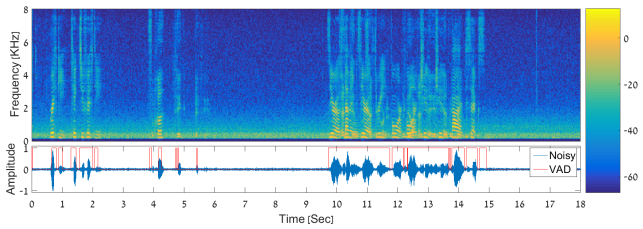
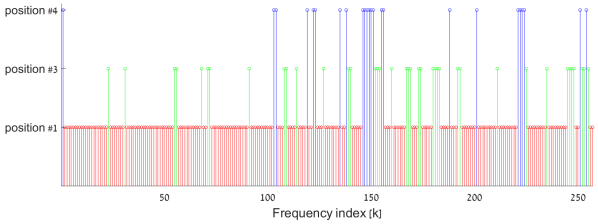
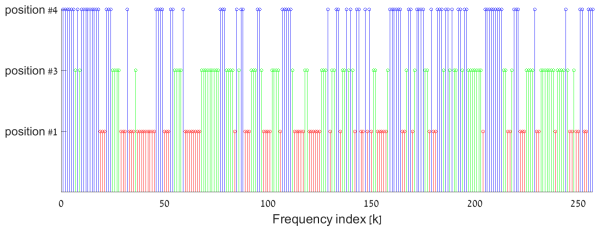



Fig. 2: Noisy signal at microphone #5, selected as the reference microphone.




(a) Only desired speaker is active.




(b) Desired and interfering speakers are jointly active.

Fig. 3: SPI decisions per frequency.

signal. Note, that the VAD accurately tracks the speech-active frames. Additionally, it can be easily verified that when both the desired and the interfering speakers are active, the VAD is on as well. In order to evaluate the proposed SPI scheme, we estimated the RTF based on time-frames classified by the VAD as speech-active. Define $I(l,k) = \text{argmax}_i\, D_i(l,k)$ the frequency-wise speaker position identifier. Figure 3 depicts these decisions for a specific set of speech time-frames. We first analyzed frames from the segment $[0 \div 3]$sec in which the desired speaker is active. Fig. 3a illustrates the frequency-wise SPI decisions. It is clear that, in this case, most frequencies are associated with position #1, which is occupied by the desired speaker with only low percentage of misclassification. The aggregated measure in (20) will therefore identify the first position as the source of the estimated RTF. We further examined time-frames where both speakers are active. It can be easily deduced from Fig. 3b, that in this segment $I(l,k)$ is not dominated by any position and hence will not be determined as either desired or interference speaker, and consequently no valid RTF can be estimated.

Finally, the estimated desired signal, $\tilde{s}(l,k)$ (16), is depicted in Fig. 4. First, the BF suppresses the interfering speaker power by approximately 20dB. The NN-MM algorithm was then applied to attenuate the residual background noise. It is evident that the background noise was significantly suppressed by the joint application of the BF and the postfilter.

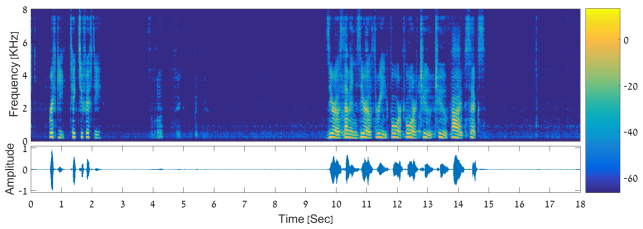



Fig. 4: The estimated signal after the proposed algorithm $\tilde{s}$.

## V. Conclusion

In this paper, a system for speaker extraction and noise reduction was presented. New SPP-based VAD controls the noise covariance matrix update, and an SPI method, which is based on an RTF library, controls the RTFs-matrix update. The updated LCMV-BF was then applied to enhance the speech. The NN-MM algorithm was used as a postfilter to attenuate the residual noise. The proposed algorithm was examined using real-life recordings in a low-reverberant enclosure, and proved to perform well in a wide range of SNR and SIR levels.

## References

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

[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. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, 1997.

[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] 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.

[6] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Transactions on Vehicular Technology*, vol. 42, no. 4, pp. 514–518, 1993.

[7] H. Q. Dam, S. Y. Low, S. Nordholm, and H. H. Dam, "Adaptive microphone array with noise statistics updates," in *Proceedings of the 2004 International Symposium on Circuits and Systems (ISCAS)*, vol. 3. IEEE, 2004, pp. 433–436.

[8] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: an analytical evaluation," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 241–252, 1999.

[9] 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.

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

[11] S. E. Chazan, S. Gannot, and J. Goldberger, "A phoneme-based pre-training approach for deep neural network with application to speech enhancement," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.

[12] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A study on manifolds of acoustic responses," in *International Conference on Latent Variable Analysis and Signal Separation (LVA-ICA)*. Liberec, Czech Republic: Springer, 2015, pp. 203–210.

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

ISBN 978-0-9928626-7-1 © EURASIP 2017