Successive Relative Transfer Function Identification Using Single Microphone Speech Enhancement

Dani Cherkassky, Shlomo E. Chazan, Jacob Goldberger, and Sharon Gannot
Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel
Email: cherkad@biu.ac.il, Shlomi.Chazan@biu.ac.il, Jacob.Goldberger@biu.ac.il, Sharon.Gannot@biu.ac.il

***Abstract*—A distortionless speech extraction in a reverberant environment can be achieved by an application of a beamforming algorithm, provided that the relative transfer functions (RTFs) of the sources and the covariance matrix of the noise are known. In this contribution, we consider the RTF identification challenge in a multi-source scenario. We propose a successive RTF identification (SRI), based on a sole assumption that sources become successively active. The proposed algorithm identifies the RTF of the ith speech source assuming that the RTFs of all other sources in the environment and the power spectral density (PSD) matrix of the noise were previously estimated. The proposed RTF identification algorithm is based on the neural network Mix-Max (NN-MM) single microphone speech enhancement algorithm, followed by a least-squares (LS) system identification method. The proposed RTF estimation algorithm is validated by simulation.**

I. Introduction

Beamforming is one of the most commonly used techniques in microphone array processing. Typically, a beamformer is used to obtain a spatial focusing on the desired speech source, while reducing the interfering sources and the background noise [1], [2]. A well-known beamforming criterion is the linearly constrained minimum variance (LCMV) aiming at minimizing the noise power at the beamformer output, under a set of linear constraints that control the array beam shape, such that the desired signal remains undistorted while interfering signals are rejected [3], [4]. In a reverberant environment, the LCMV constraints set is often expressed in terms of the relative transfer functions (RTFs) of the sources, where each RTF describes the coupling between the microphones as a response to a given source [5].

The RTF estimation challenge in a noisy environment with a single active speech source is well studied in the literature. Gannot et al. [5] exploit the nonstationarity of speech signals to estimate the RTF. Cohen [6] utilized the speech presence probability (SPP) in the time-frequency domain to identify the time-frequency instances that consist of speech signal. The time-frequency instances that consist of speech signal are then utilized to derive an RTF estimator. A subspace-based approach to RTF identification was proposed by Markovich-Golan et al. [4], where the RTF estimate is obtained by solving a generalized eigenvalue problem. A comparative survey of the covariance subtraction and the covariance whitening methods for RTF estimation was presented by Markovich-Golan and Gannot [7].

RTFs estimation in a multiple and concurrent speakers scenario was recently considered in the literature. It was proved by Hadad et al. [8] that knowledge of a basis that spans the subspace of the desired sources and a basis that spans the subspace of the interfering sources suffices for implementing the LCMV beamforming algorithm. The aforementioned desired and interfering sources subspaces can be estimated in a scenario where all the desired sources are simultaneously active and all the interfering sources are simultaneously active. However, signal segments in which desired and interfering sources are simultaneously active cannot be used for estimating the subspaces. Hassani et al. [9] proposed a method for estimating the desired and the interfering sources subspaces by exploiting signal segments with concurrent activity of the desired and the interfering sources. It was assumed that an initial estimate of the desired and interfering sources subspaces is available, then, the individual subspace estimates were projected onto the joint signal subspace of all the desired and interfering sources. The procedure results in an improved estimate of the individual subspaces as compared with the initial estimates by exploiting signal segments with concurrent activity of the desired and the interfering sources.

Deleforge et al. [10] proposed a generalization of the RTFs definition to several sources. The generalized RTFs are defined through a multichannel, multi-frame spectrograms of the received, noise-free signal.

In this work we consider a multi-source scenario. We propose a successive RTF identification (SRI) technique, based on the sole assumption that sources become successively active. Namely, we address the challenge of estimating the RTF of the ith speech source while assuming that the RTFs of all other currently active sources in the environment and the power spectral density (PSD) matrix of the noise were previously estimated. The proposed SRI algorithm first blocks the speech signals from the previously estimated RTFs. A single channel speech enhancement algorithm is then applied to one of the blocked signals and results in an estimation of the ith speech component. The estimation of the ith speech signal is then utilized as an input to a least-squares (LS) system identification algorithm which results in an estimate of the desired RTF. In this work, we utilize the NN-MM single microphone speech enhancement algorithm [11]. However, alternative single microphone speech enhancement methods are also applicable to the challenge at hand.

ISBN 978-0-9928626-7-1

II. Problem Formulation

A. Data model

Consider an array consisting of M microphones capturing a time-varying acoustical scene. Each of the involved signals propagates through the acoustic environment before being picked up by the array. In the short-time Fourier transform (STFT) domain, the nth speech source is denoted $s_n(\ell,k)$, the acoustic transfer function (ATF) relating the nth source and the mth microphone is denoted $g_{m,n}(k)$, and the stationary noise at the mth microphone is denoted $v_m(\ell,k)$, where ℓ is the frame index, and k is the frequency index. The received signals in the STFT domain can be formulated in a vector representation

$$\mathbf{z}(\ell,k)=\sum_{n=1}^{N}\mathcal{I}_n(\ell)\mathbf{h}_n(k)x_n(\ell,k)+\mathbf{v}(\ell,k), \tag{1}$$

where N is the number of sources of interest, $x_n(\ell,k)=g_{1,n}(k)s_n(\ell,k)$, $\mathcal{I}_n(\ell)\in\{0,1\}$ indicates the activity of $s_n(\ell,k)$ and $\mathbf{h}_n(k)$ is the RTF vector of the nth source defined as

$$\mathbf{h}_n(k)=\left[1,\frac{g_{2,n}(k)}{g_{1,n}(k)},\cdots,\frac{g_{M,n}(k)}{g_{1,n}(k)}\right]^T. \tag{2}$$

Considering the sources activity pattern, we assume that the speech sources become active in a successive manner. Accordingly, the activity indicator function of the nth source is defined by

$$\mathcal{I}_n(\ell)=\begin{cases}0, & \text{if } \ell\le\ell_n\\ 1, & \text{if } \ell_n<\ell\le\ell_{n+1}\\ \mathcal{A}_n, & \text{otherwise.}\end{cases} \tag{3}$$

where $\mathcal{A}_n\in\{0,1\}$ is a Bernoulli random variable. The noise $\mathbf{v}(\ell,k)$ is assumed active through the entire measurement period. The proposed activity pattern dictates that the speech sources do not become simultaneously active and that they are active for a sufficient amount of time before they become inactive again. The considered activity pattern may be practical, for example, in a noisy conference call scenario.

B. Multichannel speech extraction

In many applications, a group of the speech sources is desired while the other sources are regarded as interference. Thus, we are often interested in extracting the desired signals from the noisy measurements. The extraction can be accomplished by applying a beamformer $\mathbf{w}(\ell,k)$ to the received signal. Assuming $M>N$, $\mathbf{w}(\ell,k)$ can be chosen to satisfy the LCMV criterion [12]

$$\mathbf{w}(\ell,k)=\underset{\mathbf{w}}{\operatorname{argmin}}\left\{\mathbf{w}^H(\ell,k)\mathbf{\Phi}_{\text{vv}}(k)\mathbf{w}(\ell,k)\right\}\\ \text{subject to } \mathbf{H}^H(\ell,k)\mathbf{w}(\ell,k)=\mathbf{g}(\ell,k), \tag{4}$$

where $\mathbf{\Phi}_{\text{vv}}(k)$ is the PSD matrix of the noise $\mathbf{v}(\ell,k)$, $\mathbf{H}(\ell,k)\in\mathbb{C}^{M\times N}$ such that the nth column of $\mathbf{H}(\ell,k)$ is equal to $\mathbf{h}_n(k)$ and $\mathbf{g}(\ell,k)\in\mathbb{C}^{N\times 1}$ is the constraint vector. The well-known solution to (4) is given by

$$\mathbf{w}_{\text{LCMV}}(\ell,k)=\mathbf{\Phi}_{\text{vv}}^{-1}(k)\mathbf{H}(\ell,k)\times\\ \left(\mathbf{H}^H(\ell,k)\mathbf{\Phi}_{\text{vv}}^{-1}(k)\mathbf{H}(\ell,k)\right)^{-1}\mathbf{g}(\ell,k). \tag{5}$$

The effectiveness of $\mathbf{w}_{\text{LCMV}}$ in the desired speech extraction task is well-established [4]. In order to apply (5), one is required to estimate the RTFs matrix $\mathbf{H}(\ell,k)$ and the noise PSD matrix $\mathbf{\Phi}_{\text{vv}}(k)$. The noise PSD matrix can be estimated straightforwardly by an application of a sample covariance estimator

$$\widehat{\mathbf{\Phi}}_{\text{vv}}(k)=\frac{1}{\ell_1}\sum_{\ell=1}^{\ell_1}\mathbf{z}(\ell,k)\mathbf{z}^H(\ell,k). \tag{6}$$

The challenge of estimating the RTF of the first speech source $x_1(\ell,k)$ is well-studied in the literature, provided it is active alone. For instance, $\mathbf{h}_1(k)$ can be estimated by applying the subspace-based RTF estimator [4], which is based on the generalized eigenvalue problem for the matrix pencil $(\widehat{\mathbf{\Phi}}_{\text{zz}}(k),\widehat{\mathbf{\Phi}}_{\text{vv}}(k))$, explicitly

$$\widehat{\mathbf{\Phi}}_{\text{zz}}(k)\mathbf{u}(k)=\lambda(k)\widehat{\mathbf{\Phi}}_{\text{vv}}(k)\mathbf{u}(k), \tag{7}$$

where $\lambda(k)$ and $\mathbf{u}(k)$ is an eigenvalue-eigenvector pair and $\widehat{\mathbf{\Phi}}_{\text{zz}}(k)$ is the sample covariance of the measurements

$$\widehat{\mathbf{\Phi}}_{\text{zz}}(k)=\frac{1}{\ell_2-\ell_1}\sum_{\ell=\ell_1+1}^{\ell_2}\mathbf{z}(\ell,k)\mathbf{z}^H(\ell,k). \tag{8}$$

Due to the single-source scenario, the eigenvector $\mathbf{u}(k)$ that belongs to the largest eigenvalue $\lambda(k)$ is a scaled version of $\mathbf{h}_1(k)$. Since, by definition, the first entry of $\mathbf{h}_1(k)$ is equal to 1, the eigenvector $\mathbf{u}(k)$ can be normalized to yield an estimate of $\mathbf{h}_1(k)$

$$\widehat{\mathbf{h}}_1(k)=\frac{\widehat{\mathbf{\Phi}}_{\text{vv}}(k)\mathbf{u}(k)}{\mathbf{i}^T\widehat{\mathbf{\Phi}}_{\text{vv}}(k)\mathbf{u}(k)}, \tag{9}$$

where $\mathbf{i}=[1,0,\cdots,0]^T$. In the following section we will propose an identification procedure for the other RTFs, namely $\mathbf{h}_n(k), n>1$.

III. Successive RTF identification

When multiple speech sources are concurrently active the RTF estimators proposed in [5],[4] are not valid. In the sequel we propose the SRI algorithm for $\mathbf{h}_i(k)$ identification, under the assumption that the RTFs $\mathbf{h}_n(k), n<i$ of the previously active sources in the environment were already identified.

A. Blocking the previously estimated RTFs

Let us consider frames $\ell>\ell_i$, and assume that the estimators $\widehat{\mathbf{h}}_n(k), 0<n<i$ are already available. The received signal (1) can be projected onto the null subspace of $\left\{\widehat{\mathbf{h}}_n(k)\right\}_{n=1}^{i-1}$ by an application of the blocking matrix [13]

$$\mathbf{B}(k)=\mathbf{I}_{M\times M}-\mathbf{C}(k)(\mathbf{C}^H(k)\mathbf{C}(k))^{-1}\mathbf{C}^H(k), \tag{10}$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

where $\mathbf{C}(k) \in \mathbb{C}^{M\times i-1}$ such that the nth column of $\mathbf{C}(k)$ is equal to $\widehat{\mathbf{h}}_n(k)$. Applying the blocking matrix to the received signals results in $\mathbf{z}_b(\ell,k) = \mathbf{B}(k)\mathbf{z}(\ell,k)$:

$$\mathbf{z}_b(\ell,k) = \mathbf{B}(k)\mathbf{h}_i(k)x_i(\ell,k) + \mathbf{B}(k)\mathbf{v}(\ell,k) + \boldsymbol{\epsilon}(\ell,k), \quad (11)$$

where $\boldsymbol{\epsilon}(\ell,k) = \sum_{n=1}^{i-1}\mathbf{B}(k)\mathbf{h}_n(k)x_n(\ell,k)$. Note that under the assumption of $\widehat{\mathbf{h}}_n(k) \approx \mathbf{h}_n(k), 0 < n < i$ the additive contributions of the blocked speech sources to (11) are negligible. Consequently, we can assume that $\boldsymbol{\epsilon}(\ell,k) \approx 0$.

B. Single channel speech enhancement by NN-MM

We will now apply the NN-MM algorithm to enhance the speech component in the blocked signals (11). The NN-MM algorithm merges the generative Mixture of Gaussians (MoG) model and the discriminative deep neural network (DNN) approach. The utilized NN-MM algorithm comprises an existing phoneme-based MoG in which each Gaussian represents a different phoneme, and an existing DNN phoneme-classifier which classifies time-frame features to one of the phonemes in the phoneme-based MoG [11].

We apply the NN-MM algorithm to enhance the speech component in $z_{\mathrm{b}}(\ell,k)$[1], one of blocked signals:

$$z_{\mathrm{b}}(\ell,k) = \mathbf{i}^T\mathbf{z}_{\mathrm{b}}(\ell,k) = \alpha(k)x_i(\ell,k) + v_{\mathrm{b}}(\ell,k), \quad (12)$$

where $\alpha(k) = \mathbf{i}^T\mathbf{B}(k)\mathbf{h}_i(k)$ and $v_{\mathrm{b}}(\ell,k) = \mathbf{i}^T\mathbf{B}(k)\mathbf{v}(\ell,k)$. The noise $v_{\mathrm{b}}(\ell,k)$ is modeled by a single Gaussian, estimated during speech absence frames. The DNN estimates the phonemes probabilities, and an SPP $\rho(\ell,k)$ is calculated by applying the maximization approximation approach [14].

A soft spectral attenuation which was found useful for speech enhancement [11], [15] is then applied:

$$\begin{aligned}\tilde{x}_i(\ell,k) =& \widehat{\alpha(k)x_i(\ell,k)} \\ &= z_b(\ell,k)\rho(\ell,k) + \beta z_b(\ell,k)(1-\rho(\ell,k)) \quad (13)\end{aligned}$$

where β is a design parameter controlling the tradeoff between noise attenuation and speech distortion. In this work, $\tilde{x}_i(\ell,k)$ is utilized only for estimating $\mathbf{h}_i(k)$. Thus, noise attenuation is of higher importance than the speech distortion. Accordingly, we set β to result in an aggressive noise attenuation.

C. Least squares RTF identification

Given $\tilde{x}_i(\ell,k)$, a scaled estimate of $x_i(\ell,k)$, we can define the following LS optimization problem

$$\hat{\boldsymbol{\theta}}(k) = \underset{\boldsymbol{\theta}(k)}{\operatorname{argmin}}\|\mathbf{z}(\ell,k) - \boldsymbol{\theta}(k)\tilde{x}_i(\ell,k)\|^2. \quad (14)$$

[1]Note that an adjustment of the STFT analysis frame length may be required. The phoneme classifier requires that the frame duration of $z_{\mathrm{b}}(\ell,k)$ is equal to a typical phoneme pronunciation time. An application of the rest of the SRI algorithm dictates a frame duration which is longer than the length of the associated acoustical impulse responses in the considered enclosure. Accordingly, one may need to adjust the $z_{\mathrm{b}}(\ell,k)$ frame size prior to applying the NN-MM, and subsequently to readjust the NN-MM output signal frame size.

Algorithm 1: Sequential RTF estimation

Initialization:
1. Utilize frames $0 < \ell \le \ell_1$ to compute $\widehat{\boldsymbol{\Phi}}_{\mathrm{vv}}(k)$ using (6).
2. Utilize frames $\ell_1 < \ell \le \ell_2$ to compute $\widehat{\mathbf{h}}_1(k)$ using (9).

Upon activation of $s_i(\ell,k)$**,** $i > 1$**:**
1. Compute the blocking matrix $\mathbf{B}(k)$ using (10).
2. Compute the blocked signal $z_{\mathrm{b}}(\ell,k)$ using (11). and (12).
3. Compute $\tilde{x}_i(\ell,k)$ using NN-MM (13).
4. Compute a scaled RTF estimate $\widehat{\boldsymbol{\theta}}(k)$ using (15).
5. Normalize $\widehat{\boldsymbol{\theta}}(k)$ using (16) to result in $\widehat{\mathbf{h}}_i(k)$.

Output: $\widehat{\mathbf{h}}_i(k)$

the solution of (14) is given by [16]

$$\widehat{\boldsymbol{\theta}}(k) = \frac{\sum_{\ell=\ell_i+1}^{\ell_{i+1}}\tilde{x}_i^*(\ell,k)\mathbf{z}(\ell,k)}{\sum_{\ell=\ell_i+1}^{\ell_{i+1}}\tilde{x}_i(\ell,k)\tilde{x}_i^*(\ell,k)}. \quad (15)$$

Accordingly, we claim that $\widehat{\boldsymbol{\theta}}(k)$ is a scaled version of $\mathbf{h}_i(k)$, i.e. $\widehat{\boldsymbol{\theta}}(k) \approx \mathbf{h}_i(k)/\alpha(k)$. Since, by definition, the first entry of $\mathbf{h}_i(k)$ is equal to 1, the estimator $\widehat{\boldsymbol{\theta}}(k)$ can be normalized to yield an estimate of $\mathbf{h}_i(k)$

$$\widehat{\mathbf{h}}_i(k) = \frac{\widehat{\boldsymbol{\theta}}(k)}{\mathbf{i}^T\widehat{\boldsymbol{\theta}}(k)}. \quad (16)$$

The proposed SRI procedure is summarized in Algorithm 1. It is worth noting that the proposed estimator $\widehat{\mathbf{h}}_i(k)$ is, in general, sub-optimal since the residual noise in $\tilde{x}_i(\ell,k)$ contains components contributed by $x_j(\ell,k)\ \forall i \neq j$ and by $v_{\mathrm{b}}(\ell,k)$. Accordingly, the residual noise in $\tilde{x}_i(\ell,k)$ is correlated with $\mathbf{z}(l,k)$ and may result in biased estimate of $\widehat{\mathbf{h}}_i(k)$.

D. Practical considerations

The proposed SRI procedure, as summarized in Algorithm 1, assumes that the activity indicator function of the ith source $\mathcal{I}_i(\ell)$ is available to the algorithm. The SRI procedure utilizes $\mathcal{I}_i(\ell)$ to address the challenge induced by a *birth* of a speaker. In a practical scenario, $\mathcal{I}_i(\ell)$ should be deduced form the measurements $\mathbf{z}(l,k)$. In addition, an RTF *death* mechanisms is also required. Refer to [17] for an equivalent discussion in dynamic scenarios.

Source counting methods [18] might be useful for detecting the number of active sources in a specific time period. Since a simultaneous *birth* and *death* of two independent speakers seldom occurs, an SRI process might be triggered when an increase in the number of active sources occurs. An RTF *death* mechanism might be triggered when a decrease in the number of active speakers occurs. For example, the ith RTF may be considered as obsolete if the energy of $\hat{x}_i(\ell,k)$ is below a threshold for a predetermined period of time.

ISBN 978-0-9928626-7-1 © EURASIP 2017

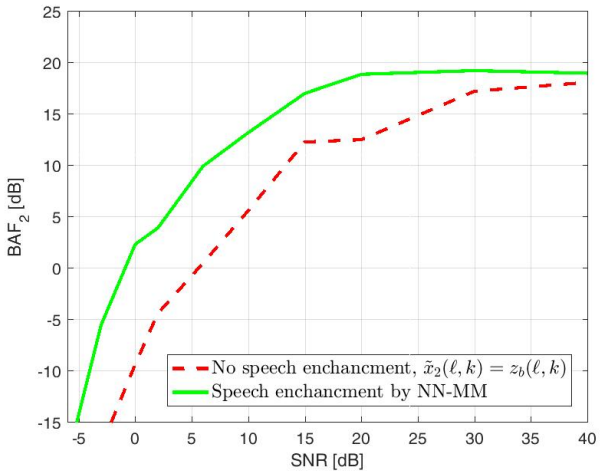

Fig. 1. Blocking ability factor.

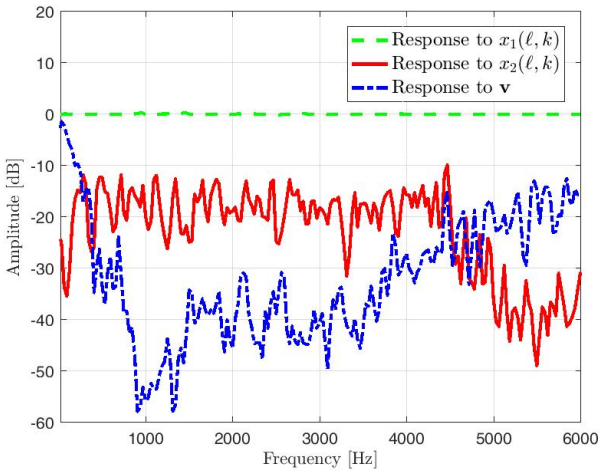

Fig. 2. $\mathbf{w}_{\text{LCMV}_1}$ responses towards the sources of interest.

IV. Simulation Study

In this section we evaluate the performance of the proposed algorithm, in a simulated $6 \times 6 \times 2.4$ m room with $T_{60} \approx 350$ mSec. A uniform linear array comprising $M = 8$ microphones with 5 cm inter-spacing was positioned in the center of the room. The sampling frequency of the system is set to 16 KHz. Three acoustic sources were positioned at a distance of 3 m from the array center. Namely, two equi-power speech sources $s_1(t)$ and $s_2(t)$ impinged on the array from angles of arrivals equal to 60° and 90°, respectively, and a stationary fan noise $v(t)$ impinged the array with an angel of arrivals equal to 120°. The powers of the sources are defined as σ_{s}^2 and σ_{v}^2, respectively. The activity pattern of the sources is set such that the time difference between $s_1(t)$ and $s_2(t)$ activation is 10 seconds, following $s_2(t)$ activation both sources remain active for additional 10 seconds. The processing is executed in the frequency domain, the STFT analysis window length is set to 512 samples for the NN-MM and to 2048 for the rest of the SRI algorithm, with 75% overlap between successive frames. We utilize the signal to noise ratio (SNR), SNR $= \sigma_{\text{s}}^2/\sigma_{\text{v}}^2$, and the blocking ability factor (BAF) to characterize the estimation results

$$\text{BAF}_n \triangleq \frac{1}{M-1}\sum_{m=2}^{M}\frac{\sigma_{m,n}^2}{\sigma_{m,\text{v}}^2}\times \frac{E\left\{\left[v_m(t)-\widehat{h}_{m,n}(t)*v_1(t)\right]^2\right\}}{E\left\{\left[x_{m,n}(t)-\widehat{h}_{m,n}(t)*x_{1,n}(t)\right]^2\right\}}, \tag{17}$$

where $x_{m,n}(t)$ is the speech generated by $s_n(t)$ as measured by the mth microphone, $v_m(t)$ is the noise at the mth microphone, $\sigma_{m,n}^2$ is the power of $x_{m,n}(t)$, $\sigma_{m,\text{v}}^2$ is the power of $v_m(t)$, $\widehat{h}_{m,n}(t)$ is the estimated RTF relating the first and the mth microphone as a response to $s_n(t)$ and $E\{[\cdot]^2\}$ is the power of $[\cdot]$. The blocking ability factor BAF_n measures the ratio between the ability to block the nth speech source and its inherent ability to block a random noise. BAF has a major effect on the amount of distortion introduced by the RTF-based GSC structure, due to desired speech leakage [5].

A. Results

In the first experiment, we tested the performance of the SRI algorithm, for various SNR values. Our main goal in this experiment is to estimate the RTF of the second source $\mathbf{h}_2(k)$, while utilizing only the time frames where both speech sources are active $\ell_2 < \ell$. In order to accomplish this task, we computed $\widehat{\mathbf{\Phi}}_{\text{vv}}(k)$ and $\widehat{\mathbf{h}}_1(k)$ by utilizing earlier time frames, as suggested in Algorithm 1. The resulting BAF of $\widehat{\mathbf{h}}_2(k)$ is depicted in Fig. 1. We compare BAF_2 for two cases. In the first case, no single-channel speech enchantment is used i.e. $\tilde{x}_2(\ell,k) = z_b(\ell,k)$, while in the second case, (13) with β set to attenuate the noise by 20 dB is applied. As can be readily observed the application of the NN-MM algorithm improves the resulting BAF_2, especially for $5 < \text{SNR} < 25$ dB.

In the second experiment, we set the SNR to 10 dB and applied Algorithm 1 to compute $\widehat{\mathbf{\Phi}}_{\text{vv}}(k)$, $\widehat{\mathbf{h}}_1(k)$ and $\widehat{\mathbf{h}}_2(k)$. Based on the estimated quantities, we computed two different LCMV beamformers using (5). The constrains set of the first beamformer $\mathbf{w}_{\text{LCMV}_1}$, is set to impose a distortionless response to a source with an RTF equal to $\widehat{\mathbf{h}}_1(k)$ and a null to a source with an RTF equal to $\widehat{\mathbf{h}}_2(k)$. The response of the second beamformer $\mathbf{w}_{\text{LCMV}_2}$, is set to be distortionless to a source with an RTF equal to $\widehat{\mathbf{h}}_2(k)$ and a null to a source with an RTF equal to $\widehat{\mathbf{h}}_1(k)$. The response to the stationary noise is unconstrained in both beamformers. The resulting responses of $\mathbf{w}_{\text{LCMV}_1}$ towards $x_1(\ell,k), x_2(\ell,k)$ and $\mathbf{v}(\ell,k)$ are depicted in Fig. 2, while the responses of $\mathbf{w}_{\text{LCMV}_2}$ towards the aforementioned sources are presented Fig. 3. As can be observed, applying $\mathbf{w}_{\text{LCMV}_1}$ to $\mathbf{z}(\ell,k)$ results in an

ISBN 978-0-9928626-7-1 © EURASIP 2017

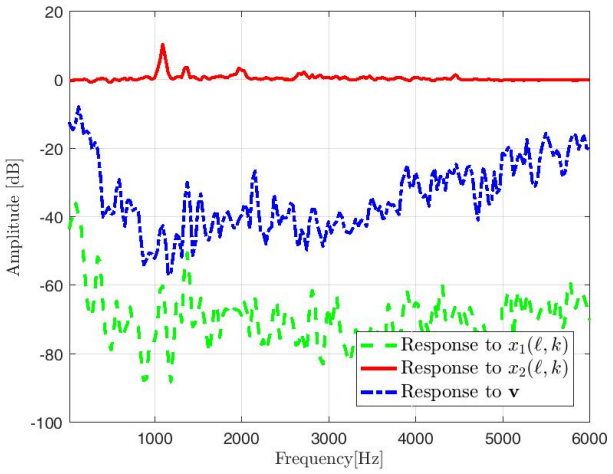

Fig. 3. $\mathbf{w}_{\text{LCMV}_2}$ responses towards the sources of interest.

attenuation of the $x_2(\ell,k)$ component by more than 10dB in the entire frequency band, while $\mathbf{w}_{\text{LCMV}_2}$ maintains a distortionless response towards $x_2(\ell,k)$ reasonably well.

V. Summary

In this paper we addressed a successive RTFs identification challenge in a multi-source scenario. We propose the SRI technique based on the assumption that sources become successively active. Particularly, we addressed the challenge of estimating the RTF of the ith speech source while assuming that the RTFs of all the other active sources in the environment were previously estimated. The proposed SRI algorithm first blocks the speech signals from the previously estimated RTFs. A single channel speech enhancement algorithm is then applied to one of the blocked signals and results in an estimation of the ith speech component. The estimation of the ith speech signal is then utilized as an input to a least-squares (LS) system identification algorithm which results in an estimate of the desired RTF. The proposed SRI method was verified in a simulative study and shown to perform well in a wide range of SNR levels.

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Berlin, Germany Springer-Verlag, 2008.

[2] S. Gannot and I. Cohen, "Adaptive beamforming and postfitering," in *Springer Handbook of Speech Processing*. New York: Springer, 2007.

[3] 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.

[4] S. Markovich-Golan, 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] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[7] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 544–548.

[8] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 543–558, 2016.

[9] A. Hassani, A. Bertrand, and M. Moonen, "LCMV beamforming with subspace projection for multi-speaker speech enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 91–95.

[10] A. Deleforge, S. Gannot, and W. Kellermann, "Towards a generalization of relative transfer functions to more than one source," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO), Nice, France*, 2015.

[11] 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.

[12] H. L. Van Trees, *Detection, estimation, and modulation theory, optimum array processing*. Wiley, New York, 2004.

[13] G. Strang, *Introduction to linear algebra*. Wellesley-Cambridge Press Wellesley, MA, 1993, vol. 3.

[14] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.

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

[16] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2007.

[17] S. Markovich-Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 201–204.

[18] O. Walter, L. Drude, and R. Haeb-Umbach, "Source counting in speech mixtures by nonparametric Bayesian estimation of an infinite gaussian mixture model," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 459–463.

ISBN 978-0-9928626-7-1