METHODS TO LEARN BANK OF FILTERS STEERING NULLS TOWARD POTENTIAL POSITIONS OF A TARGET SOURCE

Jiří Málek[1], *David Botka*[1], *Zbyněk Koldovský*[1] *and Sharon Gannot*[2]

[1]Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec,
Studentská 2, 461 17 Liberec, Czech Republic. E-mail: jiri.malek@tul.cz
[2]Faculty of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel.

ABSTRACT

In signal enhancement applications, a reference signal which provides information about interferences and noise is desired. It can be obtained via a multichannel filter that performs a spatial null in the target position, a so-called target-cancelation filter. The filter must adapt to the target position, which is difficult when noise is active. When the target location is confined to a small area, a solution could be based on preparing a bank of target-cancelation filters for potential positions of the target. In this paper, we propose two methods to learn such banks from noise-free recordings. We show by experiments that learned banks have practical advantages compared to banks that were prepared manually by collecting filters for selected positions.

Index Terms— Relative transfer functions, generalized sidelobe canceler, target-cancelation filters, noise extraction, semi-blind source separation

1. INTRODUCTION

In audio signal processing applications, such as speech enhancement, the goal is to separate a desired signal from other interferences and noise. Information about "what should be removed" from the observed signal mixture to accomplish that task is the fundamental requirement. Using multiple microphones, a multichannel filter that performs a spatial null in the target position can be designed. Signals coming from other positions can be observed at its output, which gives the key reference signal. This filter will be referred to as target-cancelation filter (CF).

Classical beamformers [1] such as the minimum variance distortionless response beamformer (MVDR), implemented as a generalized sidelobe canceler (GSC), or its improvements derived later [2], assume knowledge of whether a target-only, interferer-only or noise-only period occurs. If the answer is positive, the corresponding parameters of the approaches, such as coefficients of the CF, can be adjusted according to the observed signals. As the parameters must be kept frozen during periods when more sources are active, these methods fail when a change occurs that requires a prompt adjustment of parameters. For example, the CF providing noise reference signal must be adapted when the target changes its location.

The general aim therefore is to also adapt to changes during noisy intervals. The most popular approach is to use Blind Source Separation (BSS) [3]. If possible, BSS is strengthened by a priori knowledge, which is commonly referred to as semi-blind separation. For example, Kellermann et al. aimed at estimating the relative transfer function (RTF) between two microphones to build up a CF through a geometrically constrained BSS algorithm [4, 5]. An unbiased estimator of the RTF allowing the presence of a stationary noise was proposed in [2]; see also [6].

A new concept using a bank of CFs (CFB) where the filters are prepared for potential positions of the target in advance has recently been considered [7]; see also related works [8, 9]. It is assumed that the target's location is confined to a small area and that this area can be covered by the CFB. This means that, for any position of the target within the area, the CFB contains an efficient CF. The problem of acquiring the CF is thus reduced to selecting the best CF from the bank that maximally cancels the target [10]. The latter task is simpler than the former one; nevertheless, it is still challenging, especially when signal-to-noise ratio goes below 0 dB. A variant of the GSC using CFB called Informed Generalized Sidelobe Canceler (IGSC) has been proposed in [11].

There are two ways to obtain the CFB for a given scenario. One way, referred to as *manual*, is to collect CFs for selected positions within the assumed area [7]. These positions are typically chosen to form a grid. The other *learning* way is to gather CFs by scanning a noise-free recording of the target source that is moving within the area. Both approaches have different pros and cons. For example, the former could produce CFs that are rarely used in practice while some CFs could be missing due to directionality of the target. On the other hand, the learning approach could produce several CFs for the same position.

We address the mentioned issues in this paper. After some

This work was supported by the Czech Science Foundation through Project No. 14-11898S and by the Student Grant Scheme (SGS) at the Technical University of Liberec.

definitions that are given in Section 2, we propose two methods to learn CFBs in Section 3. Section 4 is dedicated to an experimental study of the manual and learning approaches in distant and close target speaker scenarios. Based on the results, we draw some conclusions, emphasizing the benefits of the learning approach.

2. TARGET-CANCELLATION FILTERS

A two-channel[1] noisy recording of a target whose position is fixed is described through

$$x_{\rm L}(n) = \{h_{\rm L} * s\}(n) + y_{\rm L}(n), \tag{1}$$

$$x_{\rm R}(n) = \{h_{\rm R} * s\}(n) + y_{\rm R}(n), \tag{2}$$

where n is the time index, $*$ denotes the convolution, $x_{\rm L}$ and $x_{\rm R}$ are, respectively, the signals from the left and right microphone, s is the target signal, and $y_{\rm L}$ and $y_{\rm R}$ are the other signals, hereinafter simply referred to as noise. $h_{\rm L}$ and $h_{\rm R}$ denote the microphone-target impulse responses that depend on the target's position and on the acoustical environment.

Alternatively, we can interpret (1) as

$$x_{\rm L}(n) = s_{\rm L}(n) + y_{\rm L}(n), \tag{3}$$

$$x_{\rm R}(n) = \{h_{\rm rel} * s_{\rm L}\}(n) + y_{\rm R}(n), \tag{4}$$

where $s_{\rm L} = h_{\rm L} * s$ and $h_{\rm rel}$ is the relative impulse response between the microphones. The goal of the enhancement is to retrieve either $s_{\rm L}$ or $s_{\rm R} = h_{\rm R} * s$ (dereverberation is not the goal here). The frequency-domain description reads

$$X_{\rm L}(\omega) = S_{\rm L}(\omega) + Y_{\rm L}(\omega), \tag{5}$$

$$X_{\rm R}(\omega) = H_{\rm rel}(\omega) S_{\rm L}(\omega) + Y_{\rm R}(\omega), \tag{6}$$

where $H_{\rm rel}(\omega)$ is the relative transfer function (RTF) between the channels, that is, the Fourier transform of $h_{\rm rel}$. Since $H_{\rm rel}(\omega) = H_{\rm R}(\omega)/H_{\rm L}(\omega)$, it is sometimes also called the transfer function ratio [2, 12].

A CF consists of two filters $g_{\rm L}$ and $g_{\rm R}$, and its output is $g_{\rm L} * x_{\rm L} - g_{\rm R} * x_{\rm R}$. The output should be free of the target signal s. The filters $g_{\rm L}$ and $g_{\rm R}$ can be defined in many ways – see, e.g., [10, 12, 15] – but the most popular option, which will be considered here, is $g_{\rm L} = h_{\rm rel}$ and $g_{\rm R}(n) = \delta(n - D)$ (the delayed unit impulse), or, in the frequency domain, $G_{\rm L}(\theta) = H_{\rm rel}(\theta)$ and $G_{\rm R}(\theta) = e^{-iD\theta}$. In this case, a CF is determined only by $g_{\rm L}$ and by the integer delay D, which can be equal to a constant.

Since $g_{\rm L}$ depends on $h_{\rm L}$ and $h_{\rm R}$, it must be estimated from data recorded on-site. When a noise-free recording of the target is available ($y_{\rm L} = y_{\rm R} = 0$), the filter can be estimated from the recording using least squares estimation [10, 14] or frequency-domain estimates [2, 12].

2.1. Performance criteria

To evaluate the cancelation performance of a CF or of the entire CFB on a block of signals, we measure the output variance of the filter or of the filter bank, respectively, when applied to the noiseless target signal. It follows that the smaller the output variance[2], the better the target cancelation is.

For a CF given by $g_{\rm L}$, the criterion is

$$\texttt{ovar}(g_{\rm L}) = \frac{1}{M} \sum_n v^2(n), \tag{7}$$

where $v = x_{\rm L} * g_{\rm L} - x_{\rm R} * g_{\rm R}$ is the output of the CF and the sum goes over a block of data of length M. For filter bank A, the criterion is

$$\texttt{ovar}(\mathrm{A}) = \min_{i \in \{1,\dots,|\mathrm{A}|\}} \texttt{ovar}(g_{\rm L}^i), \tag{8}$$

where $g_{\rm L}^i$ is the ith filter in A, and $|\mathrm{A}|$ denotes the number of filters in A. Here, it is assumed that the best canceling CF from the bank is, at any given moment, the one yielding the minimum output variance. Note that this assumption holds when the recording is free of noise.

For the entire recording, the performance of a bank A can be evaluated as $\texttt{ovar}(\mathrm{A})$ averaged over all blocks of the recording. We will denote this criterion as $\overline{\texttt{ovar}}(\mathrm{A})$.

The output variance of a filter is proportional to the scale of the input signals. To avoid this dependence, we always normalize the input signals to unit variance (averaged over both channels) before evaluating $\overline{\texttt{ovar}}(\mathrm{A})$.

3. METHODS TO LEARN THE CFB

Assume that a noise-free recording of a target source is available. During the recording, the target is moving randomly within the confined area. We propose two learning algorithms to derive efficient filter banks that cover the positions visited during the recording. Note that a secondary objective is to minimize the number of filters in the bank.

3.1. One-pass CFB training

The method described here scans the noise-free recording in one pass, so it might be useful in real-time scenarios. The data are processed block by block, and the blocks may overlap.

In the beginning, the bank is empty. The algorithm computes CF in the first block and puts it into the bank if its output variance is smaller than the threshold denoted by τ_1. If the variance is higher, the algorithm proceeds in the next block and this is repeated until a first CF is added to the bank.

Then, each block is, in parallel, filtered by all CFs in the current bank and their outputs are compared. Let the ith CF $g_{\rm L}^i$ be the filter yielding the minimum output variance. If

[1] For simplicity, we consider only two channels, but the ideas of this paper can be generalized to more channels.

[2] See [11] for other criteria performing well also when noise is present.

Algorithm 1: The one-pass learning algorithm.

```
Input: noise-free training signals $x_{\mathrm{L}}$ and $x_{\mathrm{R}}$
Output: L1-CFB
L1-CFB=$\emptyset$;$K = 0$
Initialization: find the first CF whose ovar is smaller than $\tau_1$
foreach block of $x_{\mathrm{L}}$ and $x_{\mathrm{R}}$ do
    $i = \arg\min_{k\in\{1,\dots,K\}} \mathtt{ovar}(g_{\mathrm{L}}^k)$;
    if $\mathtt{ovar}(g_{\mathrm{L}}^i) > \tau_1$ then
        compute $g_{\mathrm{L}}$ from the block being processed;
        if $\mathtt{ovar}(g_{\mathrm{L}}) < \tau_2 \cdot \mathtt{ovar}(g_{\mathrm{L}}^i)$ then
            $g_{\mathrm{L}}^{K+1} = g_{\mathrm{L}}$;
            L1-CFB=L1-CFB $\cup\{g_{\mathrm{L}}^{K+1}\}$;
            $K = K + 1$;
        end
    end
end
```

Algorithm 2: The two-pass learning method.

```
Input: L1-CFB, the number of filters $N$
Output: L2-CFB
L2-CFB = L1-CFB; $K = $|L2-CFB|;
while $K > N$ do
    for $i \in \{1, \dots, K\}$ do
        A = L2-CFB $\setminus\{g_{\mathrm{L}}^i\}$;        /* $g_{\mathrm{L}}^i$ is excluded */
        $\Delta_i = \overline{\mathtt{ovar}}(\mathrm{A}) - \overline{\mathtt{ovar}}(\mathrm{L2\text{-}CFB})$;
    end
    $k = \arg\min_{i\in\{1,\dots,K\}} \Delta_i$;
    L2-CFB = L2-CFB $\setminus\{g_{\mathrm{L}}^k\}$;
    $K = K - 1$;
end
```

$\mathtt{ovar}(g_{\mathrm{L}}^i)$ is higher than τ_1, a new CF is computed for the block of signals. The resulting filter is added to the bank if its output variance is lower than $\tau_2 \cdot \mathtt{ovar}(g_{\mathrm{L}}^i)$ where $\tau_2 < 1$. We use $\tau_2 \approx 0.79$, which corresponds to 1 dB improvement compared to $\mathtt{ovar}(g_{\mathrm{L}}^i)$. The procedure is described in Algorithm 1. The resulting bank will henceforth be referred to as L1-CFB.

The computational complexity of this method is mainly due to the computation of new CFs, while the burden due to the other operations is negligible. In the worst case, a CF has to be computed in each block. A real-time run is thus possible provided that the method to compute the CF is fast enough.

3.2. CFB training with verification

In an off-line regime, it is possible to verify the cancelation performance of the CFB and withdraw those CFs that seem to be redundant. We propose a method that allows the user to select the number of filters in the resulting bank. The method starts from L1-CFB. Then, it successively excludes CFs whose removal causes minimum deterioration of $\overline{\mathtt{ovar}}$ until the required number of filters is achieved. The method is described in Algorithm 2.

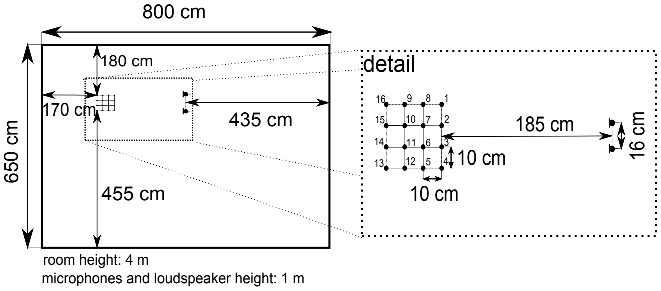


Fig. 1. Room of the SISEC 2013 dataset

4. EXPERIMENTS

The following experiments are designed to compare trained banks (L-CFB) with manually constructed banks (M-CFB). An M-CFB consists of CFs each of which is computed from 1 s of training noise-free recording of the target (a loudspeaker) standing in a fixed position. The L-CFBs are computed from 60 s long training signals (moving the loudspeaker or microphones) using a 1 s analyzing window with a shift length of 125 ms.

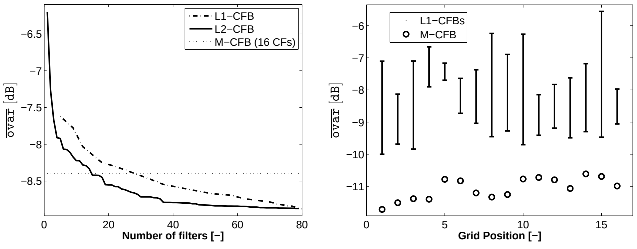
We consider two scenarios illustrated in Figures 1 and 3. The first one concerns a situation where the target-microphone distance is large, about 2 m, while the other is focused on small distances, from 10 to 20 cm. Banks are constructed so that the manual banks comprise CFs for positions denoted by numbered points. The learned banks are trained on recordings of a source moving smoothly along all of these points. All recordings are sampled at 16 kHz. The length of each computed CF is $1,000$ and the delay parameter $D = 20$.

For testing, other recordings of a source moving randomly within the considered area are used. The recordings are normalized to unit variance due to the performance evaluation (see Section 2.1). The evaluation proceeds on blocks of length 2000 samples. The criterion $\overline{\mathtt{ovar}}$ is used to compare the banks.

4.1. Distant speaker

This scenario resembles a situation where the target is a speaker sitting in a noisy meeting room and its position is limited due to the movements of the speaker's head. The dataset of the SISEC 2013 evaluation campaign [16] is used from the "Two-channel noisy recordings of a moving speaker within a limited area" task.

Signals of the dataset are recorded in a large meeting room (8×6.5 m) with reverberation time $T_{60} \approx 650$ ms. The target is a loudspeaker placed within a 30×30 cm area. It is always directed towards two microphones that are 2 m distant from the center of the area. The M-CFB consists of 16 CFs for the marked positions in the detail of Fig. 1. The length of the test recording is 30 s.



(a) (b)

Fig. 2. Results for the distant-speaker task of SISEC 2013

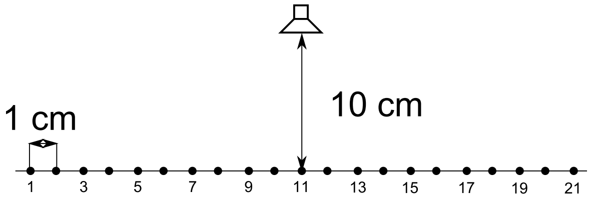


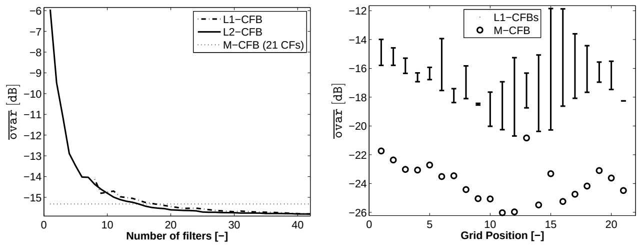
Fig. 3. Sketch of the scenario with mobile phone model

Fig. 2(a) shows $\overline{\texttt{ovar}}$ of the computed banks when the number of CFs[3] varies from 1 through 80. The performance of the banks improves with an increasing number of filters. The dotted line shows the performance of M-CFB comprised of a fixed number of 16 CFs. These results show that the L1-CFB and the M-CFB are equivalent when the former consists of 28 CFs (the M-CFB always consists of 16 filters). The L2-CFB demonstrates equivalent performance with merely 12 CFs and is hence advantageous.

Performance degradation of the L-CFBs algorithms with respect to the M-CFB is always exhibited for the nominal points of the latter. Indeed, the M-CFB is optimized for the nominal points while L-CFBs contain filters for random positions that are close to these points. Fig. 2(b) demonstrates this by assessing the banks on recordings of the target source from positions 1 through 16. The circles show the $\overline{\texttt{ovar}}$ of M-CFB for each position, while the bars show the range of $\overline{\texttt{ovar}}$ of L1-CFB and L2-CFB when the number of contained CFs varies from 1 through 80. The losses in terms of $\overline{\texttt{ovar}}$ range from 1 dB to 6 dB, which is acceptable in many practical situations.

4.2. Close speaker

In [10], a noise reduction method using a bank of CFs for the case when a person is speaking into a cell phone is proposed. In this application, the position of the speaker is mostly limited to the immediate vicinity of the phone. To imitate such situation, we use a mobile phone model from [10] and an artificial head. The target, which is a loudspeaker inside the head, is standing in a fixed position while the phone mock-up is moving. The area of movements is linear as shown in Fig 3. The experiment was realized in a 3.5×4 m room having 2 m in height and $T_{60} \approx 300$ ms.



(a) (b)

Fig. 4. Results of experiments in scenario where the speaker is close to microphones

The M-CFB comprises 21 filters for linearly arranged positions. The L-CFBs were trained on recording during which the model was moved roughly along the line (± 1.5cm close to the line). The length of the test utterance was 16 s.

Results of the experiments are shown in Figures 4(a) and 4(b). They are in correspondence with the results of Section 4.1 but they also show that the practical conclusions could be different for the close-speaker scenario. Namely, L1-CFB and L2-CFB are equivalent to the M-CFB with 17 and 15 CFs, respectively, while the M-CFB consists of 21 CFs. Both learning methods require a lower number of CFs to achieve the same $\overline{\texttt{ovar}}$ as the M-CFB. Although L2-CFB is still better than L1-CFB, the differences are small compared to the distant-speaker scenario; cf. Fig. 2(a) and Fig. 4(a). It is therefore more beneficial to use L1-CFB in this case, due to its simpler construction.

The performance degradation of the L-CFB algorithms compared to M-CFB in nominal positions 1 through 21 ranges between 2 and 10 dB. This loss is higher compared to the distant-speaker experiment (c.f. Fig. 2(b) and Fig. 4(b)). Nevertheless, since the average performance is by about 10 dB better in this close-speaker scenario, the loss of up to 10 dB could be affordable.

5. CONCLUSIONS

Without claims of generalizing the presented results, it is seen that the learned banks of CFs have important advantages compared to their manual counterparts. In particular, they can adapt to an irregular shape of the area of target occurrence or to typical movements of the target. They can be directly utilized in connection with Informed Generalized Sidelobe Canceler [11] or within the noise reduction method for mobile phones [10].

[3] In case of L1-CFB, the number of filters is driven by the threshold parameter τ_1.

6. REFERENCES

[1] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing*, John Wiley & Sons, 2004.

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[3] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, K. Shikano, "Blind Spatial Subtraction Array for Speech Enhancement in Noisy Environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 650–664, May 2009.

[4] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, W. Kellermann, "Geometrically Constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4, 2013.

[5] L. C. Parra, C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[6] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2055-2063, Aug. 1996.

[7] Z. Koldovský, J. Málek, P. Tichavský, and F. Nesta, "Semi-blind Noise Extraction Using Partially Known Position of the Target Source," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 21, no. 10, pp. 2029-2041, Oct. 2013.

[8] D. Model and M. Zibulevsky, "Signal reconstruction in sensor arrays using sparse representations," *Signal Processing*, vol. 86, no. 3, pp. 624–638, March 2006.

[9] R. Talmon and S. Gannot, "Relative transfer function identification on manifolds for supervised GSC beamformers," in *Proc. of the 21st European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.

[10] Z. Koldovský, P. Tichavský, D. Botka, "Noise Reduction in Dual-Microphone Mobile Phones Using A Bank of Pre-Measured Target-Cancellation Filters," *Proc. of ICASSP 2013*, pp. 679–683, Vancouver, Canada, May 2013.

[11] J. Málek, Z. Koldovský, S. Gannot, and P. Tichavský, "Informed Generalized Sidelobe Canceler Utilizing Sparsity of Speech Signals," *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2013)*, Southampton, UK, Sept. 2013.

[12] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, Jan. 2011.

[13] Z. Koldovský and P. Tichavský, "Time-domain blind separation of audio sources on the basis of a complete ICA decomposition of an observation space", *IEEE Trans. on Speech, Audio and Language Processing*, vol. 19, no. 2, pp. 406–416, Feb. 2011.

[14] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Information Theory*, vol. 40, no. 2, pp. 340-349, 1994.

[15] Y. Lin, J. Chen, Y. Kim and D. Lee, "Blind channel identification for speech dereverberation using ℓ_1 norm sparse learning," *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, MIT Press, Vancouver, British Columbia, Canada, December 3-6, 2007.

[16] N. Ono, Z. Koldovský, S. Miyabe, N. Ito, "The 2013 Signal Separation Evaluation Campaign," *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, Sept. 2013. [online] `http://sisec.wiki.irisa.fr`.