MULTICHANNEL SPEECH ENHANCEMENT USING CONVOLUTIVE TRANSFER FUNCTION APPROXIMATION IN REVERBERANT ENVIRONMENTS

Ronen Talmon[1], *Israel Cohen*[1] *and Sharon Gannot*[2]

[1] Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
{ronenta2@tx, icohen@ee}.technion.ac.il

[2] School of Engineering
Bar-Ilan University
Ramat-Gan, 52900, Israel
gannot@eng.biu.ac.il

ABSTRACT

Recently, we have presented a transfer-function generalized sidelobe canceler (TF-GSC) beamformer in the short time Fourier transform domain, which relies on a convolutive transfer function approximation of relative transfer functions between distinct sensors. In this paper, we combine a delay-and-sum beamformer with the TF-GSC structure in order to suppress the speech signal reflections captured at the sensors in reverberant environments. We demonstrate the performance of the proposed beamformer and compare it with the TF-GSC. We show that the proposed algorithm enables suppression of reverberations and further noise reduction compared with the TF-GSC beamformer.

Index Terms— Adaptive signal processing, array signal processing, acoustic noise, speech enhancement, speech dereverberation

1. INTRODUCTION

Multi-channel speech enhancement algorithms using microphone arrays have been an active area of research for many years and are known to have the potential of improving single sensor solutions. In reverberant environments, the signal acquired by a microphone array is distorted by the acoustic impulse response and usually corrupted by noise. Beamforming techniques, which aim at recovering the desired source signal from the reverberant and noisy measurements, are the most common and studied solutions.

Linearly constrained minimum variance (LCMV) adaptive beamforming, proposed by Frost [1], and in particular its generalized sidelobe canceler (GSC) unconstrained version developed by Griffiths and Jim [2] are the most commonly used beamforming methods. These methods assume that the signals captured at the sensors are delayed versions of the source signal. However in reverberant environments, the speech signal is propagated through room impulse response. Gannot *et al*. [3] proposed the so-called transfer function GSC (TF-GSC), which exploits the acoustic path by incorporating the relative transfer function (RTF) to achieve noise reduction. This approach is carried out in the short time Fourier transform (STFT) domain and approximates the linear convolution as a multiplicative transfer function (MTF). Recently, we have presented a GSC framework in the STFT domain using a complete representation of a linear convolution [4], and proposed a new practical algorithm relying on the convolutive transfer function approximation (CTF-GSC). Performance evaluation showed that the CTF-GSC is especially advantageous in reverberant environments and achieves both improved noise reduction and reduced speech distortion. In addition, it was shown that the improved performance is obtained mainly due to the use of an improved RTF identification method [5]. It is worthwhile noting that the main objective of both the TF-GSC and the CTF-GSC is noise reduction, i.e. recovering the reverberant speech component captured at one of the sensors.

In this paper, we incorporate a delay-and-sum beamformer into the GSC structure under the CTF approximation presented in [4]. This beamformer is designed to steer the beam towards a single direction of the desired source location, while minimizing the response in the reflections and noise source directions. We show that in reverberant environments the proposed approach enables suppression of the speech signal reflections along with improved noise reduction compared with the TF-GSC performance. This paper is organized as follows. In Section 2, we formulate the problem in the STFT domain. In Section 3, we present the proposed algorithm. Finally, in Section 4, we show experimental results that demonstrate the advantages of the proposed method.

2. PROBLEM FORMULATION

Consider an array of M microphones in a noisy and reverberant environment, where a single speech source located inside the enclosure. The output of the mth microphone is given by

$$\begin{aligned} y_m(n) &= a_m(n) * s(n) + u_m(n), m = 1, 2, ..., M \\ &\triangleq d_m(n) + u_m(n) \end{aligned} \quad (1)$$

where $*$ denotes convolution, $s(n)$ represents a (non-stationary) speech source, $a_m(n)$ represents the acoustic room impulse response between the speech source and the mth microphone and $d_m(n)$ and $u_m(n)$ are the speech and noise components received at the mth microphone. We assume that the noise signals $u_m(n), m = 1, 2, ..., M$ are stationary and uncorrelated with the speech source. Alternatively, the measurements can be represented with respect to the speech component at the first microphone

$$y_m(n) = h_m(n) * d_1(n) + u_m(n), m = 1, 2, ..., M \quad (2)$$

where $h_m(n)$ represents the *relative* impulse response between the mth microphone and the first microphone with respect to the speech source location, which satisfies $a_m(n) = h_m(n) * a_1(n)$.

The signals can be divided into overlapping time frames and analyzed using the short time Fourier transform. Let P denote the number of time frames of the source signal $s(n)$, N denote the length of each time frame, and L denote the framing step. According to

This research was supported by the Israel Science Foundation (grant no. 1085/05).

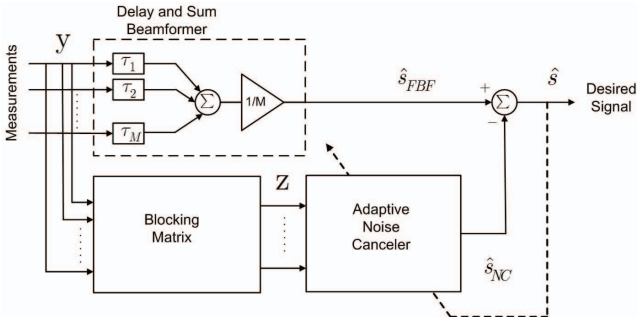

Fig. 1. Proposed GSC structure.

[6], [7], [8] a filter convolution in the time domain is transformed into a sum of N cross-band filter convolutions in the STFT domain. Hence, we can represent (2) in the STFT domain

$$\begin{aligned} y_m(p,k) &= d_m(p,k) + u_m(p,k), m = 1,2,...,M \\ &= \sum_{k'=0}^{N-1} h_m(p,k',k) * d_1(p,k') + u_m(p,k) \end{aligned} \quad (3)$$

where p is the time frame index, k and k' are the frequency sub-band indices and $h_m(p,k',k)$ is the cross-band filter between frequency band k' and k. Let $\mathbf{d}_m(k)$, $\mathbf{u}_m(k)$ and $\mathbf{y}_m(k)$ denote column stack vectors of length P comprised of the STFT samples at sub-band k of the signals $d_m(n)$, $u_m(n)$ and $y_m(n)$ respectively, and let $\mathbf{H}_m(k',k)$ denote the convolution matrix of the cross band filter $h_m(p,k',k)$ of size $P \times P$. Then, (3) can be written in matrix representation

$$\mathbf{y}_m(k) = \sum_{k'=0}^{N-1} \mathbf{H}_m(k',k)\,\mathbf{d}_1(k') + \mathbf{u}_m(k). \quad (4)$$

Now, by applying the convolutive transfer function (CTF) approximation [4] [5] we obtain

$$\mathbf{y}_m(k) = \mathbf{H}_m(k,k)\,\mathbf{d}_1(k) + \mathbf{u}_m(k) \quad (5)$$

where the convolution in the time domain is approximated as a convolution in each sub-band in the STFT domain.

3. PROPOSED METHOD

Based on the GSC structure in the STFT domain [4], we propose a scheme which enables reflections suppression combined with improved noise reduction. The proposed structure is similar to the GSC scheme and is comprised of three building blocks formed in two parallel processing branches, as illustrated in Fig. 1. The first block is a delay-and-sum beamformer aimed at enhancing the speech and reducing the noise. The second block is a blocking matrix which is designed to block the desired speech signal and produce noise-only outputs. The third block is a noise canceler which is built adaptively to cancel the residual noise at the fixed beamformer output given the noise-only signals.

3.1. Fixed Beamformer

The first block is a fixed delay-and-sum beamformer. In order to support broadband signals and non-integers delays, the beamformer is implemented as a filter-and-sum beamformer, i.e. applying a finite impulse response (FIR) filter to each microphone output and then summing the filtered signals [1]. The basic idea is to delay each microphone output by a proper amount of time so that the speech components from the direct path of the desired source are synchronized across all sensors. These delayed measurements are then weighted and summed. Since they add up together coherently, the speech components are reinforced. In contrast, the reflections and noise components are suppressed as they are added together destructively.

Let $\hat{s}_{FBF}(n)$ denote the output signal of the fixed beamformer (FBF) output, and let $\hat{\mathbf{s}}_{FBF}(k)$ denote a column stack vector of length P comprised of the STFT samples of $\hat{s}_{FBF}(n)$. It is worthwhile noting that $\hat{s}_{FBF}(n)$ still contains noise, originated from reflections arriving from the speech source direction.

3.2. Blocking Matrix

Let $z_m(n)$ denote the mth output signal of the blocking matrix, where $m = 2,\ldots,M$, defined as

$$z_m(n) = y_m(n) - h_m(n) * y_1(n). \quad (6)$$

Under the CTF model, (6) can be written in the STFT domain

$$\mathbf{z}_m(k) = \mathbf{y}_m(k) - \mathbf{H}_m(k,k)\mathbf{y}_1(k) \quad (7)$$

where $\mathbf{z}_m(k)$ is defined similarly to $\mathbf{y}_m(k)$. By substituting (5) into (7), we have

$$\mathbf{z}_m(k) = \mathbf{u}_m(k) - \mathbf{H}_m(k,k)\mathbf{u}_1(k).$$

Thus, the $M-1$ output signals of the blocking matrix contain noise-only components.

Implementing the blocking matrix requires estimates of the RTFs under the CTF approximation $\hat{\mathbf{H}}_m(k,k)$. Thus, we use an RTF identification method adapted to speech signals, which assumes knowledge of speech presence probabilities [5]. It is worthwhile noting that traces of the speech signal may leak into the reference noise signals due to imperfect estimation.

3.3. Noise Canceler

Let $\hat{s}_{NC}(n)$ denote the output of the noise canceler, defined as

$$\hat{s}_{NC}(n) = \sum_{m=2}^{M} g_m^{NC}(n) * z_m(n) \quad (8)$$

where $g_m^{NC}(n)$ is the noise canceler filter of the mth output signal of the blocking matrix. Under the CTF approximation, the noise canceler is reduced to a band-to-band filter at each sub-band. Thus, similarly to (5) and (7), (8) can be written in the STFT domain as

$$\hat{\mathbf{s}}_{NC}(k) = \sum_{m=2}^{M} \mathbf{G}_m^{NC}(k)\mathbf{z}_m(k) \quad (9)$$

where $\mathbf{G}_m^{NC}(k)$ is a convolution matrix of the band-to-band filter of $g_m^{NC}(n)$. In order to achieve maximal noise reduction we aim at minimizing the energy of the output signal, i.e.

$$\min_{\mathbf{G}_m^{NC}(k)} \left\|\hat{\mathbf{s}}_{FBF}(k) - \hat{\mathbf{s}}_{NC}(k)\right\|^2. \quad (10)$$

Now, the noise canceler filters $\left\{\mathbf{G}_m^{NC}(k)\right\}$ are built adaptively using the LMS algorithm.

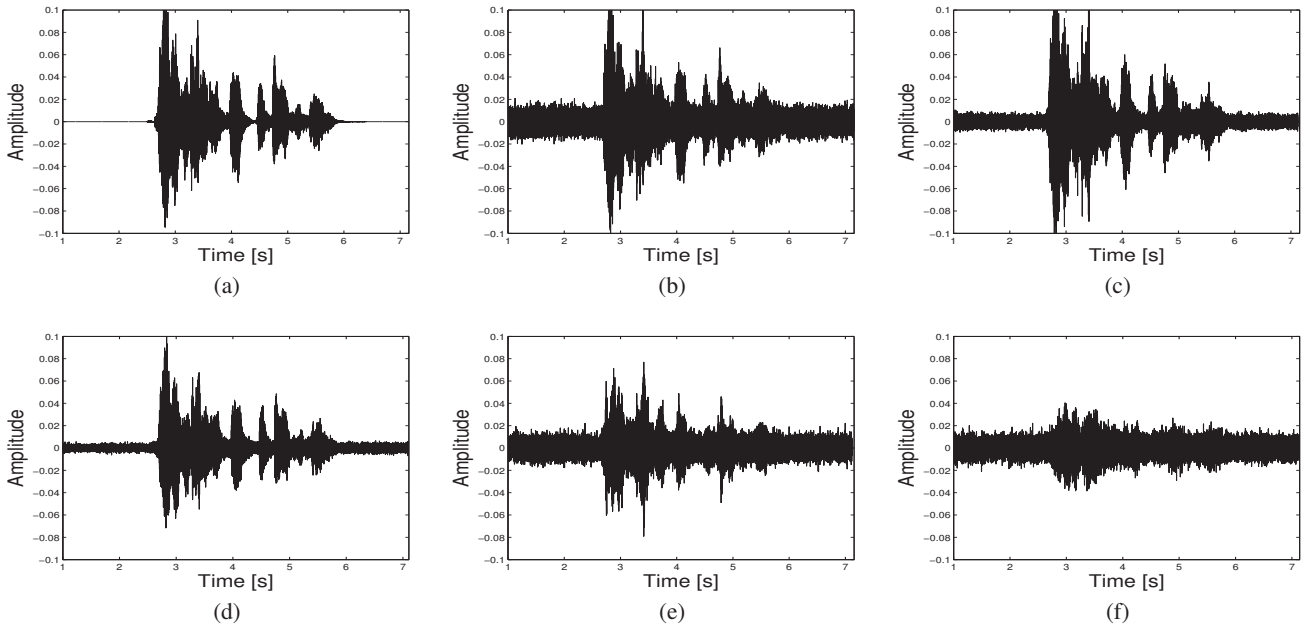

Fig. 2. Signal waveforms. (a) Reverberant speech source received at the first microphone. (b) Noisy signal received at the first microphone, SNR=5dB. (c) Enhanced signal obtained at the TF-GSC output. (d) Enhanced signal obtained at the proposed method output. (e) Reference noise signal at the output of the TF-GSC blocking matrix. (f) Reference noise signal at the end of the proposed blocking matrix.

4. EXPERIMENTAL RESULTS

In this section we evaluate the performance of the proposed method. In order to compete with a method that relies on similar assumptions, we compare the proposed method with an improved version of the TF-GSC technique. The original version of the TF-GSC, proposed in [3], is based on the *non-stationarity* RTF identification method [9], which assumes the presence of non-stationary source and stationary uncorrelated noise. We improve this algorithm by replacing the RTF identification method with a method adapted to speech signals [10] which also takes advantage of silent periods. Thus, both the improved version of the TF-GSC and the proposed method require knowledge of speech presence probabilities, which can be obtained using a voice activity detector (VAD).

In the following experiments we simulate the acoustic impulse responses according to the image method [11]. The responses are measured in a rectangular room, 4 m wide by 7 m long by 2.75 m high. We locate a linear 5 microphone array at the center of the room, at $(2, 3.5, 1.375)$. The microphone array topology consists of 5 microphones in a horizontal straight line with $(3, 5, 7, 9)$ cm spacings. The primary microphone (designated here as the "first" microphone) was set to be the microphone positioned at the middle of the array. A speech source at $(2, 5.5, 1.375)$ is 2 m distant from the primary microphone[1], and a noise source is placed at $(1.5, 4, 1.375)$.

The signals are sampled at 8 kHz. The speech source signal is a recorded speech from the TIMIT database [12] and the noise source signal is a computer generated white zero mean Gaussian noise with variance that varies to control the SNR level. The microphones measurements are generated by convolving the source signals with corresponding simulated acoustic impulse responses. We use a short period of noise-only signal at the beginning of each experiment for estimating the noise signals PSDs and for adaptive adjustment of the noise canceler. In practice, the noise PSDs can be evaluated online using a voice activity detector (VAD). The STFT is implemented using Hamming windows of length $N = 512$ with 50% overlap.

In order to compare the performance of the competing algorithms, we use the noise reduction (NR) measure, defined by

$$\text{NR} \triangleq 10 \log_{10} \frac{\sum\limits_{n \in T_n} y_1^2(n)}{\sum\limits_{n \in T_n} \hat{s}^2(n)}$$

where T_n denotes periods where the speech signal is absent.

In the first experiment, we compare the performances of the two competing methods in a reverberant environment with reverberation time set to 0.5s. Figure 2(a)-(f) shows the waveform of the speech component received by the primary microphone, the noisy measurement at the primary microphone with SNR level of 5 dB, the enhanced speech at the output of the TF-GSC and the proposed methods, and a reference noise signal obtained at the output of the blocking matrix in both methods. We observe that the noise level at the output of the proposed method is lower than the noise level at the output of the TF-GSC method. In addition, the reference noise signal at the output of the blocking matrix of the proposed method contains less traces of the speech, which yields better noise cancelation and less speech distortion at the beamformer output. Figure 3 shows the noise reduction curves of the TF-GSC and the proposed method in various input SNR conditions. The proposed method obtains significantly better noise reduction than the TF-GSC in all tested SNR levels, with constant difference.

In the second experiment, we demonstrate the reflections attenuation achieved by the proposed method. It is worthwhile noting that the TF-GSC method aims at producing undistorted reverberant

[1] Creating a far-end field configuration.

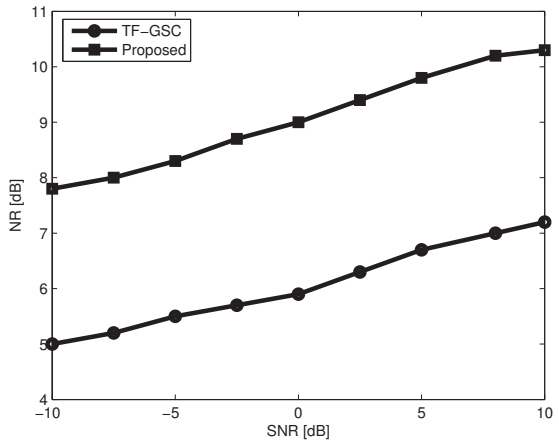

Fig. 3. Noise reduction (NR) curves under various SNR conditions.

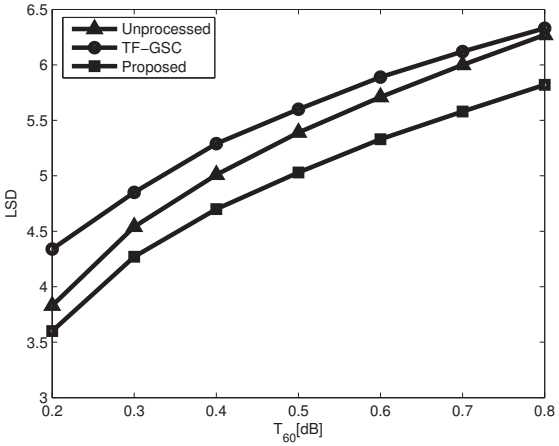

Fig. 4. Mean log spectral distance (LSD) curves for various reverberation times.

speech as captured at the primary microphone. For evaluation we use the log spectral distance (LSD) measure defined as

$$\mathrm{LSD}(p) \triangleq \left[\frac{1}{N}\sum_{k=0}^{N-1}\left|\ell(s(p,k)) - \ell(\hat{s}(p,k))\right|^2\right]^{\frac{1}{2}}$$

where

$$\ell\left(f(t)\right) = 10\log_{10}\left|f(t)\right|.$$

Figure 4 shows the mean LSD curves obtained in various reverberation times with SNR level of 10 dB. We observe that the speech component at the output of the upper branch of the proposed method is less reverberant than the unprocessed speech signal captured at the primary sensor. In addition, we obtain that the speech component at the output of the upper branch of the TF-GSC is more reverberant than the unprocessed signal, mainly due to imperfect RTF identification.

5. CONCLUSION

We have presented an improved version of the generalized sidelobe canceler (GSC) beamformer in the STFT domain under the convolutive transfer function approximation. The proposed algorithm combines a delay-and-sum beamformer with the GSC structure in order to suppress the speech signal reflections captured at the sensors in reverberant environments. We demonstrated the performance of the proposed method and compared it with the TF-GSC method. It was shown that the proposed method enables better noise reduction and that the proposed beamformer output is less reverberant than the signals captured at the sensors. It is worthwhile noting that dereverberation solutions are widely spread. Thus, incorporating more advanced dereverberation techniques into the GSC beamformer is a promising lead which requires further research.

6. REFERENCES

[1] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Jan. 1972.

[2] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly contrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

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

[4] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *submitted to IEEE Trans. Audio, Speech and Language Processing*, 2008.

[5] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *to appear in IEEE Trans. Audio, Speech and Language Processing*, 2008.

[6] Y. Avargel and I. Cohen, "System identification in the short time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[7] M.R. Portnoff, "Time frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Signal Processing*, vol. ASSP-28, no. 1, pp. 55–69, Feb. 1980.

[8] S. Farkash and S. Raz, "Linear systems in Gabor time-frequency space," *IEEE Trans. Signal Processing*, vol. 42, no. 3, pp. 611–617, Jan. 1994.

[9] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 40, no. 8, pp. 2055–2063, Aug. 1996.

[10] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech and Audio Processings*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[12] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continous speech database," National Inst. of Standards and Technology (NIST), Gaithersburg, MD, Feb. 1993.