A Distortionless Subband Beamformer for Noise Reduction in Reverberant Environments

Emanuël A.P. Habets[†]
Department of Electrical and Electronic Engineering
Imperial College London, Exhibition Road, SW7 0AT London, United Kingdom
Email: e.habets@imperial.ac.uk

***Abstract*—In this paper a distortionless subband beamformer for noise reduction in reverberant environments is described and evaluated. The subband beamformer is based on a recently proposed spatio-temporal prediction model (STPM). An alternative expression is deduced that does not require the explicit estimation of the STPM. In addition, online estimators for the speech and noise power spectral density matrices are proposed that are based on the conditional speech presence probability (SPP). Finally, the SPP is used to enhance the output of the beamformer. The presented results demonstrate the subband beamformer's ability to achieve significant noise reduction with low speech distortion.**

I. Introduction

The problem of noise reduction using a microphone array has been an active area of research for many years (see [1], [2] and the references therein). In reverberant environments, the signals acquired by the microphone array are distorted by the acoustic impulse responses and are usually contaminated by noise.

In the last decade, researchers have focused on estimating the desired reverberant signal as received by one of the microphones rather than the desired anechoic signal [1], [3]. In case of the minimum variance distortionless response (MVDR) beamformer, it was recently shown that the largest amount of noise reduction is achieved when no attempt is made to reduce reverberation [4].

Gannot et al. [3] derived a transfer function generalized sidelobe canceler (TF-GSC) to estimate the desired signal as received by a reference microphone. This is achieved by utilizing relative transfer functions between a reference microphone and other microphones with respect to the desired source. In [5], this beamformer was derived in the short-time Fourier transform (STFT) domain. Unlike the formulation in [3], which is based on the multiplicative transfer function (MTF) approximation, the MVDR beamformer in [5] is derived using a convolutive transfer function (CTF) approximation. The CTF approximation, which was shown to be more accurate and less restrictive, enables representations of long acoustic impulse responses with short time frames. Estimation techniques of the relative transfer functions were proposed using the MTF approximation [3], [6] and using the CTF approximation [7].

Especially in highly reverberant environments long filters are required to achieve high noise reduction and low speech distortion. In addition, it is well known that the ability of most beamformers to reduce certain types of noise such as sensor and diffuse ambient noise is limited. Therefore, a post-filter is often used in conjunction with the beamformer. The post-filter is usually formulated in the subband domain and operates on subband signals that are obtained by analyzing time frames in the order of 32 ms.

There are several advantages of also implementing the beamformer in the subband domain. Firstly, it increases the flexibility of exchanging information between the beamformer and the post-filter and hence allows for a seamless integration of linear and non-linear filters. Secondly, low latency can be achieved which is paramount for applications such as hearing-aids or hands-free communication.

In this paper, an optimal subband beamformer is described and evaluated that is based on a recently proposed spatio-temporal prediction model (STPM) [1]. Furthermore, an alternative expression for this beamformer is deduced that only depends on the power spectral density (PSD) matrices of the desired source and noise. Online estimator for the PSD matrices are proposed that is based on the speech presence probability (SPP). Finally, the beamformer is evaluated in enclosures with different reverberation times.

The paper is organised as follows. In Section II the problem is formulated. In Section III the optimal subband beamformer is described. In Section IV the estimators for the PSD matrices are described. Finally, the performance of the beamformer is evaluated and conclusions are provided in Sections VI and VII, respectively.

II. Problem Formulation

In this work, a microphone array is considered that is placed in a noisy and reverberant environment in which one desired speech source is located. The m-th microphone signal at discrete time n can be expressed as

$$y_m(n) = a_m(n) * x_1(n) + v_m(n), \quad m = 1, 2, \ldots, M, \tag{1}$$

where $a_m(n)$ denotes the relative impulse response between the m-th microphone and the first microphone with respect to the desired source location, $x_1(n)$ the desired reverberant signal received at the first microphone, and $v_m(n)$ the noise received at the m-th microphone. In the following it is assumed that the noise is uncorrelated with the desired speech signal. It is worthwhile noting that $a_m(n)$ for $m = 2, 3, \ldots, M$ is generally of infinite length [3], [5]. As the energy of the relative impulse response decays rapidly, the assumption that the support of $a_m(n)$ is finite is practically not very restrictive.

The signals can be divided into overlapping time frames and analyzed using the STFT. Let N denote the length of each time frame, and R denote the framing step. To avoid distortions due to circular convolution, each time frame is zero padded with N samples before taking the discrete Fourier transform. According to [8] a filter convolution in the time domain is transformed into a sum of cross-band filter convolutions in the STFT domain. The cross-band filters are used for cancelling the aliasing caused by sampling in each frequency subband [9]. As in [5], the CTF approximation is applied such that the m-th microphone signal in time frame ℓ and subband k can be expressed as

$$y_m(\ell, k) = \mathbf{A}_m(k)\mathbf{x}_1(\ell, k) + v_m(\ell, k), \tag{2}$$

where $\mathbf{A}_m(k)$ denotes the convolution matrix that contains the band-to-band filters for the k-th subband corresponding to relative impulse response $a_m(n)$, $\mathbf{x}_1(\ell, k)$ denotes the STFT samples of the

[†] This research was supported by the Marie Curie Intra European Fellowship within the 7th European Community Framework Programme under contract number PIEF-GA-2009-237246.

reverberant signal at the first microphone and $v_m(\ell, k)$ denotes the STFT samples of the noise at the m-th microphone.

In the sequel we assume that L STFT samples of every microphone signal are used per subband to compute the beamformer. The output of the beamformer, denoted by z, is then given by[1]

$$z = \mathbf{h}^H\,\mathbf{y}, \tag{3}$$

where $(\cdot)^H$ denotes the Hermitian transpose operator, $\mathbf{h} = [\mathbf{h}_1^T, \mathbf{h}_2^T, \dots, \mathbf{h}_M^T]^T$ with $\mathbf{h}_m = [h_m(0), h_m(1), \dots, h_m(L-1)]^T$ and $\mathbf{y} = [\mathbf{y}_1^T, \mathbf{y}_2^T, \dots, \mathbf{y}_M^T]^T$ with $\mathbf{y}_m(\ell) = [y_m(\ell), y_m(\ell-1), \dots, y_m(\ell-L+1)]^T$. The time domain signal $z(n)$ can be computed efficiently using a weighted overlap-add technique.

The objective of this work is to estimate an undistorted and noise-free version of the reverberant speech component received at the first microphone, i.e., estimate $x_1(n)$ given the microphone signals $\{y_m(n)\}_{m=1}^{M}$.

III. Distortionless Subband Beamformer

In this section an optimal subband beamformer is described. First, the STPM is discussed as proposed in [1, p. 95]. Secondly, we describe a subband MVDR beamformer that utilizes the STPM.

A. Spatio-temporal prediction

In general, the speech distortion of the reverberant signal received by the first microphone that is caused by the filter is given by

$$e_x = (\mathbf{h} - \mathbf{c})^H\,\mathbf{x}, \tag{4}$$

where $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_M^T]^T$ with $\mathbf{x}_m(\ell) = [x_m(\ell), x_m(\ell-1), \dots, x_m(\ell-L+1)]^T$ and $\mathbf{c} = [\mathbf{u}, 0, \dots, 0]^T$ is a column vector of length ML with $\mathbf{u} = [1, 0, \dots, 0]^T$ of length L. In [1] the authors proposed to use a STPM to predict the reverberant signals at microphones $m = 2, 3, \dots, M$ using the reverberant signal received at the first microphone. Under the CTF approximation, the STPM can be used to express the reverberant signal $\mathbf{x}$ in the subband domain as

$$\mathbf{x} = \mathbf{W}\mathbf{x}_1, \tag{5}$$

where

$$\mathbf{W} = [\mathbf{I}_{L\times L}, \mathbf{W}_2, \dots, \mathbf{W}_M]^T \tag{6}$$

denotes the spatio-temporal prediction matrix of size $ML \times L$ and $\mathbf{I}_{L\times L}$ is a identity matrix of size $L \times L$. By substituting (5) in (4) we find that the speech distortion is equal to

$$e_x = \left(\mathbf{W}^H\mathbf{h} - \mathbf{u}\right)^H \mathbf{x}_1, \tag{7}$$

which is zero when $\mathbf{W}^H\mathbf{h} = \mathbf{u}$. Finally, the optimal solution for $\mathbf{W}$, in the Wiener sense, can be found by minimizing

$$E\left\{(\mathbf{x} - \mathbf{W}\mathbf{x}_1)^H\,(\mathbf{x} - \mathbf{W}\mathbf{x}_1)\right\} \tag{8}$$

which leads to

$$\mathbf{W}_{\text{o}} = \boldsymbol{\Phi}_x\mathbf{C}\left(\mathbf{C}^T\boldsymbol{\Phi}_x\mathbf{C}\right)^{-1}, \tag{9}$$

where $\boldsymbol{\Phi}_x = E\{\mathbf{x}\mathbf{x}^H\}$ denotes the PSD matrix of size $ML \times ML$ of the reverberant speech signal and $\mathbf{C} = [\mathbf{I}_{L\times L}, \mathbf{0}_{L\times L}, \dots, \mathbf{0}_{L\times L}]^T$.

B. Minimum variance distortionless response beamformer

Now the aforementioned objective can be formulated in the STFT domain as

$$\mathbf{h}_{\text{STP}} = \arg\min_{\mathbf{h}} \mathbf{h}^H\boldsymbol{\Phi}_v\mathbf{h} \text{ subject to } \mathbf{W}^H\mathbf{h} = \mathbf{u}, \tag{10}$$

where $\boldsymbol{\Phi}_v = E\{\mathbf{v}\mathbf{v}^H\}$. Using Lagrange multipliers, the optimal solution for the k-th subband is obtained:

$$\mathbf{h}_{\text{STP}} = \boldsymbol{\Phi}_v^{-1}\mathbf{W}\left[\mathbf{W}^H\boldsymbol{\Phi}_v^{-1}\mathbf{W}\right]^{-1}\mathbf{u}, \tag{11}$$

By substituting (9) in (11) the following expression can be deduced

$$\mathbf{h}_{\text{STP}} = \boldsymbol{\Phi}_v^{-1}\,\boldsymbol{\Phi}_x\,\mathbf{Q}\,\mathbf{u}, \tag{12}$$

with

$$\mathbf{Q} = \mathbf{C}\left(\mathbf{C}^T\boldsymbol{\Phi}_x\boldsymbol{\Phi}_v^{-1}\boldsymbol{\Phi}_x\mathbf{C}\right)^{-1}\mathbf{C}^T\boldsymbol{\Phi}_x\mathbf{C}. \tag{13}$$

Now let us assume that $L = 1$, such that (12) can be expressed as

$$\tilde{\mathbf{h}}_{\text{STP}} = \tilde{\boldsymbol{\Phi}}_v^{-1}\,\tilde{\boldsymbol{\Phi}}_x\,\mathbf{q}, \tag{14}$$

with

$$\mathbf{q} = \tilde{\mathbf{c}}\left(\tilde{\mathbf{c}}^T\tilde{\boldsymbol{\Phi}}_x\tilde{\boldsymbol{\Phi}}_v^{-1}\tilde{\boldsymbol{\Phi}}_x\tilde{\mathbf{c}}\right)^{-1}\tilde{\mathbf{c}}^T\tilde{\boldsymbol{\Phi}}_x\tilde{\mathbf{c}}, \tag{15}$$

where all symbols with a tilde are related to the current time frame only, i.e., $\tilde{\boldsymbol{\Phi}}_x = E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H\}$ with $\tilde{\mathbf{x}} = [x_1, x_2, \dots, x_M]^T$ is a subset of $\mathbf{x}$ and $\tilde{\mathbf{c}} = [1, 0, \dots, 0]^T$ of length M. Using the fact that the speech correction matrix is rank-one, i.e.,

$$\tilde{\boldsymbol{\Phi}}_x = \phi_s\,\tilde{\mathbf{A}}\tilde{\mathbf{A}}^H, \tag{16}$$

where ϕ_s is the PSD of the desired source signal and $\tilde{\mathbf{A}} = [\tilde{\mathbf{A}}_1, \tilde{\mathbf{A}}_2, \dots, \tilde{\mathbf{A}}_M]^T$, we obtain after some manipulations

$$\tilde{\mathbf{h}}_{\text{R1-MVDR}} = \frac{\tilde{\boldsymbol{\Phi}}_v^{-1}\,\tilde{\boldsymbol{\Phi}}_x}{\text{tr}\{\tilde{\boldsymbol{\Phi}}_v^{-1}\,\tilde{\boldsymbol{\Phi}}_x\}}\tilde{\mathbf{c}}, \tag{17}$$

which is equal to the MVDR beamformer in [1, p. 134]. When short time frames are used, the rank-one assumption might be violated and the performance of (12) (for $L = 1$) and (17) can be different.

The noise and speech PSD matrices are usually unknown and therefore need to be estimated online. In Section IV, the estimation of the PSD matrices $\boldsymbol{\Phi}_x$ and $\boldsymbol{\Phi}_v$ is discussed.

C. Discussion

As mentioned in Section II the relative transfer functions (RTFs) are non-causal. Consequently, the beamformer will always introduce an additional delay of the output signal of the beamformer. While $\mathbf{A}$ in (2) can be used as a predictor, it does not exploit the underling structure of the speech signal. Furthermore, the estimation of the STPM is relatively simple compared to the estimation of the RTFs.

Although we have used only causal STFT samples in our original problem formulation, it is worthwhile mentioning that both causal and non-causal spatio-temporal prediction filters are provided by (9). Specifically, we can use any of the L prediction filters by redefining $\mathbf{u}$ used in (11) and (12). The solution that results in the largest amount of noise reduction and lowest amount of speech distortion, depends on the array configuration and the location of the sources.

Finally, it is interesting to note that no specific assumption is made regarding the desired signal in (5). In practice, multiple desired talkers might be present (either talking one at a time or at the same time). We can therefore redefine the signal $\mathbf{x}$ such that it contains all desired talkers. Because there is nothing that dissuades us from finding a STPM for this situation, the beamformer in (11) and (12) can still satisfy (10). The latter cannot be achieved when using (17) because the rank-one assumption is violated.

[1] When possible, the time frame index ℓ and/or the subband index k is omitted for notational convenience.

IV. Estimating Speech and Noise PSD Matrices

In this section we first review a recently proposed SPP estimator. Under the speech presence uncertainty, we then propose recursive estimators for the noise and speech PSD matrices.

A. Estimating speech presence probability

Recently, a multichannel SPP estimator was derived based on a Gaussian statistical model for the STFT samples. In the multichannel case the conditional SPP in the ℓ-th time frame and k-th subband is given by [10]

$$p = \left[1 + \frac{q}{1-q}\left[1+\xi\right]\exp\left(-\frac{\zeta}{1+\xi}\right)\right]^{-1}, \tag{18}$$

where

$$\zeta = \tilde{\mathbf{y}}^H \tilde{\mathbf{\Phi}}_v^{-1} \tilde{\mathbf{\Phi}}_x \tilde{\mathbf{\Phi}}_v^{-1} \tilde{\mathbf{y}}, \tag{19}$$

$$\xi = \mathrm{tr}\{\tilde{\mathbf{\Phi}}_v^{-1}\, \tilde{\mathbf{\Phi}}_x\}, \tag{20}$$

denotes the *a priori* signal to noise ratio (SNR) and q denotes the *a priori* speech absence probability and $\tilde{\mathbf{\Phi}}_x = \tilde{\mathbf{\Phi}}_y - \tilde{\mathbf{\Phi}}_v$. The latter is important as the conditional SPP is computed under the hypothesis that speech is present.

To increase the reliability of the conditional SPP, random fluctuations of ζ and ξ are reduced by time-frequency smoothing. A global and local smoothed version of ζ and ξ, represented by $\bar{\zeta}$ and $\bar{\xi}$, is computed using

$$\bar{\zeta}(\ell,k) = \frac{1}{2\Delta_k + \Delta_\ell + 2} \sum_{l'=\ell-\Delta_\ell}^{\ell} \sum_{k'=k-\Delta_k}^{k+\Delta_k} \zeta(\ell',k') \tag{21}$$

and

$$\bar{\xi}(\ell,k) = \frac{1}{2\Delta_k + \Delta_\ell + 2} \sum_{l'=\ell-\Delta_\ell}^{\ell} \sum_{k'=k-\Delta_k}^{k+\Delta_k} \xi(\ell',k'). \tag{22}$$

Different values for Δ_k and Δ_ℓ are chosen to obtain global and the local averages. Finally, an estimate of the conditional SPP is obtained using a global SPP (p_{global}) and local SPP (p_{local}) [calculated using $\bar{\zeta}$ and $\bar{\xi}$ in (18)], i.e.,

$$\hat{p} = p_{\text{global}}\, p_{\text{local}}. \tag{23}$$

Note that prior knowledge of the spatial location of the desired source can be incorporated into (23). While the speech absence probability (SAP) was fixed in the current work, a time and frequency dependent SAP can be used (see for example [11]).

B. Estimating noise PSD matrix

In total L time frames are required to estimate the noise PSD matrix $\mathbf{\Phi}_v$. As it is paramount that no speech is present during the estimation, we first compute the probability that only noise is present in all L time frames:

$$\hat{p}_v(\ell) = \prod_{\ell'=\ell-L+1}^{\ell} \left[1 - \hat{p}(\ell')\right]. \tag{24}$$

Under speech presence uncertainty, $\hat{p}_v(\ell)$ can be employed to carry out a recursive averaging:

$$\begin{aligned}\hat{\mathbf{\Phi}}_v(\ell) = {} & \left[1-\hat{p}_v(\ell)\right]\hat{\mathbf{\Phi}}_v(\ell-1) + \\ & \hat{p}_v(\ell)\left[\alpha_v \hat{\mathbf{\Phi}}_v(\ell-1) + (1-\alpha_v)\,\mathbf{y}(\ell)\mathbf{y}^H(\ell)\right],\end{aligned} \tag{25}$$

where α_v is a predefined forgetting factor. To further reduce the risk of updating the noise PSD when speech is present, we only apply (25) when $\hat{p}_v(\ell)$ is larger than a predefined threshold T_v $(0 < T_v \leq 1)$, otherwise $\hat{\mathbf{\Phi}}_v(\ell) = \hat{\mathbf{\Phi}}_v(\ell-1)$.

C. Estimating speech PSD matrix

We can estimate the PSD matrix $\mathbf{\Phi}_x$ in a similar way as the noise PSD matrix. First the probability that speech is present is all L time frames is computed using

$$\hat{p}_x(\ell) = \prod_{\ell'=\ell-L+1}^{\ell} \hat{p}(\ell'). \tag{26}$$

Using the the relation $\mathbf{\Phi}_y = \mathbf{\Phi}_x + \mathbf{\Phi}_v$ and $\hat{p}_x(\ell)$, the speech PSD matrix can be updated recursively using

$$\begin{aligned}\hat{\mathbf{\Phi}}_x(\ell) = {} & \left[1-\hat{p}_x(\ell)\right]\hat{\mathbf{\Phi}}_x(\ell-1) + \\ & \hat{p}_x(\ell)\left[\alpha_x \hat{\mathbf{\Phi}}_x(\ell-1) + (1-\alpha_x)\left\{\hat{\mathbf{\Phi}}_y(\ell) - \hat{\mathbf{\Phi}}_v(\ell)\right\}\right],\end{aligned} \tag{27}$$

where α_x is a forgetting factor and $\hat{\mathbf{\Phi}}_y(\ell)$ denotes the PSD matrix of the received signal that is estimated recursively using a forgetting factor α_y. To reduce the risk of updating the speech PSD matrix when no speech is present, (27) is updated only when $\hat{p}_x(\ell) > T_x$, where T_x $(0 < T_x \leq 1)$ is a predefined threshold, otherwise $\hat{\mathbf{\Phi}}_x(\ell) = \hat{\mathbf{\Phi}}_x(\ell-1)$.

V. Incorporating SPP into Subband Beamfomer

The conditional SPP can be used to further enhance the noise reduction of the beamfomer. Let us define two hypotheses:

$$\begin{aligned} &\mathrm{H}_0: \quad \mathbf{y} = \mathbf{v}, && \text{speech absence;} \\ &\mathrm{H}_1: \quad \mathbf{y} = \mathbf{x} + \mathbf{v} && \text{speech presence.} \end{aligned}$$

The proposed subband beamformer is then given by

$$\mathbf{h}_{\text{STP-SPP}} = \hat{p}\,\mathbf{h}_{\mathrm{H}_1} + (1-\hat{p})\,\mathbf{h}_{\mathrm{H}_0}. \tag{28}$$

where $\mathbf{h}_{\mathrm{H}_1}$ is the desired filter under the hypothesis that speech is present and $\mathbf{h}_{\mathrm{H}_0}$ is the desired filter under the hypothesis that speech is absent.

Under hypothesis H_1 we chose $\mathbf{h}_{\mathrm{H}_1} = \mathbf{h}_{\text{STP}}$ to minimize the speech distortion. Under hypothesis H_0, maximum noise reduction is obtained by $\mathbf{h}_{\mathrm{H}_0} = 0$. Although the latter achieves maximum noise reduction, there are two major disadvantages. Firstly, the obtained beamformer modulates the residual noise in a perceptually unpleasant way. Secondly, in case of false-alarm the speech signal is severely distorted. Alternatively, we can aim at reducing the noise such that

$$E\left\{\left\|\mathbf{h}_{\mathrm{H}_0}^H\mathbf{y} - 10^{-\Lambda/20}\,v_1\right\|^2 \middle| H_0\right\} \tag{29}$$

is minimized. Therefore, $\mathbf{h}_{\mathrm{H}_0} = 10^{-\Lambda/20}\mathbf{c}$, where Λ (in dB) controls the maximum amount of noise reduction in the k-th subband under the hypothesis that speech is absent.

VI. Performance Evaluation

A scenario with one desired source and one interfering source in a homogenous and spatially white noise field was considered. A linear microphone array was used with $M = 4$ microphones and an inter-microphone distance of 5 cm. The room size was $5 \times 4 \times 6$ m (length×width×height) with a reverberation time of 450 ms. All room impulse responses were generated using an efficient implementation of the source-image method [12]. The distance between a source and the first microphone is indicated by r and the incidence angle in degrees by θ. The desired source (with $r = 1$ m and $\theta = 130^\circ$) consisted of 20 s of male and female speech. The interference source (with $r = 1.5$ m and $\theta = 50^\circ$) was USASI noise. The additive noise from the spatially white noise field is zero-mean Gaussian noise; the SNR was fixed at 30 dB. The

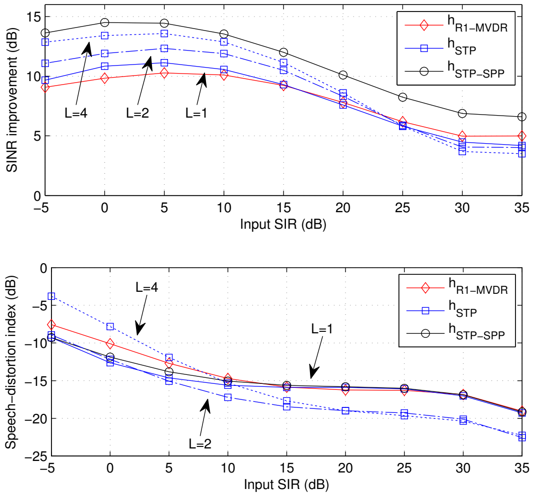


Fig. 1. SINR improvement (top) and the speech-distortion index (bottom) for different subband beamfomers and input SIRs.

parameters used were: $f_s = 8$ kHz, $N = 256$, $R = 128$, $q = 0.5$, $\alpha_y = \alpha_x = \alpha_v = 0.96$, $T_v = T_x = 0.9$, $\Delta_l = 1$, $\Delta_{\text{local},k} = 1$, $\Delta_{\text{global},k} = 3$, and $\Lambda = 20$ dB. For the first 10 time frames it was assumed that the desired source is absent and hence $q(\ell) = 1$ for $1 \leq \ell \leq 10$. No additional prior information was used to assist the beamformer.

The signal to interference plus noise ratio (SINR) improvement and the speech-distortion index (SDI) as defined in [1] were used to evaluate the performance of the subband beamformers; the first microphone was used as a reference. The results in Fig. 1 are obtained using $\mathbf{h}_{\text{R1-MVDR}}$, $\mathbf{h}_{\text{STP}}$ with $L \in \{1, 2, 4\}$ and $\mathbf{h}_{\text{STP-SPP}}$ with $L = 1$ for different input signal to interference ratios (SIRs) and 50 Monte Carlo simulations where the setup was translated and rotated within the environment. In case the input SIR is similar to the SNR the beamfomer's ability to reduce interference plus noise is limited by the noise [1], [4]. We observe that the SINR is improved and SDI is reduced for when using $L = 2$ rather than $L = 1$. For $L = 4$ we can only improve both performance measures when the input SIR is larger than 10 dB. For the considered scenario the $\mathbf{h}_{\text{STP}}$ (with and without SPP) performance is similar or better than that of the $\mathbf{h}_{\text{R1-MVDR}}$. The incorporation of the SPP into the beamformer significantly enhances the noise reduction with little increase in speech distortion.

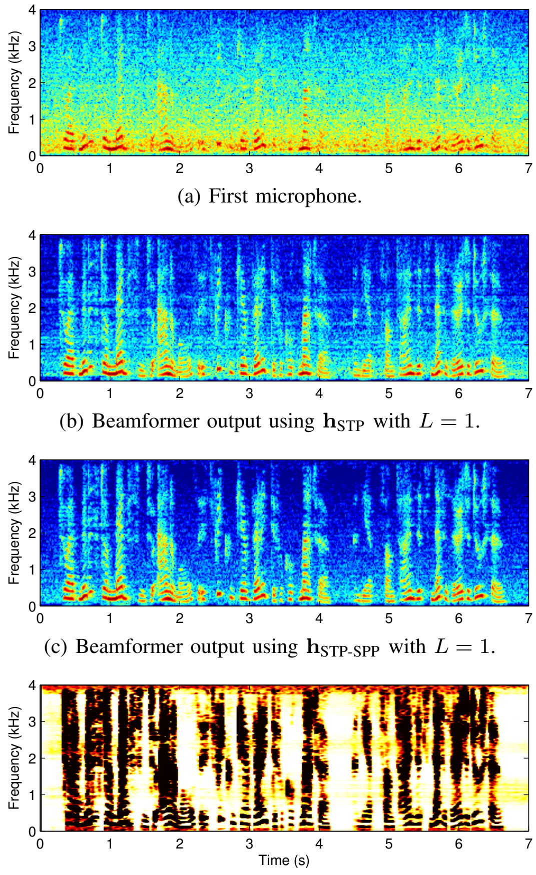
The spectrograms of the first 7 s of the first microphone signal (SIR= 5 dB and SNR= 30 dB), the output obtained using (12) and (17) as well as the estimated SPP are shown in Fig. 2. These results support the results in Fig. 1 and demonstrate that a significant amount of noise is reduced with low distortion of the desired speech. Because the SDI for $\mathbf{h}_{\text{STP-SPP}}$ is very similar to the SDI for $\mathbf{h}_{\text{STP}}$, we can conclude that the proposed conditional SPP estimator produces accurate results for the considered test conditions.

VII. Conclusion

In this work a distortionless subband beamformer was described and evaluated. In addition, online estimators were proposed to estimate the speech and noise PSD matrices using the conditional SPP. The experimental results presented in this work indicate that the use of longer subband filters can result in more noise reduction and less speech distortion in case the SIR is sufficiently high.

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.



(a) First microphone.

(b) Beamformer output using $\mathbf{h}_{\text{STP}}$ with $L = 1$.

(c) Beamformer output using $\mathbf{h}_{\text{STP-SPP}}$ with $L = 1$.

(d) Estimated speech presence probability.

Fig. 2. Spectrograms of the first 7 s of received and enhanced signals (a-c) and the speech presence probability $\hat{p}$ (black indicates speech presence and white indicates speech absence). The SIR= 5 dB and SNR= 30 dB.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. Ray Liu, Eds. Wiley, 2008, ch. 9.

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

[4] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 158–170, 2010.

[5] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Sep. 2009.

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

[7] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

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

[9] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[10] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *to appear in IEEE Trans. Audio, Speech, Lang. Process.*, 2010.

[11] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[12] E. A. P. Habets. (2008, May) Room impulse response (RIR) generator. [Online]. Available: http://home.tiscali.nl/ehabets/rirgenerator.html