INCORPORATING AUDITORY PROPERTIES INTO GENERALISED SIDELOBE CANCELLER

Adam Borowicz and Alexandr Petrovsky

Bialystok University of Technology
Department of Digital Media and Computer Graphics
Wiejska str. 45A, Bialystok, Poland
email: borowicz@wi.pb.edu.pl

ABSTRACT

A novel speech enhancement method based on generalized sidelobe canceller (GSC) and auditory properties is presented. We show that it is possible to reduce audible speech distortions and preserve residual noise level under system model uncertainties. It can be done by constraining a speech leakage power according to masking effect phenomena. An optimal noise cancellation filter is derived using constrained minimization of the residual noise power. We implemented the GSC structure using a simple delay-and-sum beamformer and corresponding (delays-dependent) blocking matrix. Finally a comparative evaluation of the selected methods is performed using objective speech quality measures. The results show that the novel method outperforms conventional one providing lower speech distortions and comparable noise attenuation.

Index Terms— GSC, speech enhancement, beamforming

1. INTRODUCTION

Speech enhancement has been an active area of research for many years. It arises in a wide range of speech processing applications including mobile radio devices, speech coding, speech recognition systems, aids for the hearing impaired. A traditional objective of the speech enhancement is to reduce environmental noise while preserving speech intelligibility. In a context of the multichannel systems the dereverberation and interference suppression is also expected.

Over the past decades most efforts have been devoted to the dereverberation and beamforming techniques. The key idea of the beamforming is to process the microphone array signals to listen the sounds coming from only one direction. Particularly the noise reduction can be implicitly achieved by avoiding 'noisy' directions. The linearly constrained minimum variance (LCMV) algorithm has been originally proposed by Frost [1] in the 1970s and it is probably the most studied beamforming method since then. It minimizes beamformer output variance subject to the set of linear equations that ensure a constant gain in a specified listening direction. The minimum variance distortionless response (MVDR) method [2] can be considered as a special case of the LCVM approach. Another popular technique is generalized sidelobe canceller [3]. The noisy signal domain is split into two orthogonal subspaces where the dereverberation and noise suppression can be performed separately. Unfortunately due to model uncertainties a speech signal leakages to the noise subspace which results in increased speech distortions. There are also other approaches to multichannel speech enhancement [4], but they don't consider the dereverberation problem, and try to recover only reverberant noise-free microphone signal, thus they are out of scope of this paper.

The proposed system is based on the conventional GSC beamformer. However we directly assume the presence of the system model uncertainties and use masking effects to reduce speech leakage effect. A similar technique has been proved to be useful in several single channel methods [5] but is rarely used in a field of the multichannel speech enhancement. It is observed, that for a given spectral power level, there is a masking threshold so that any interferer below this threshold becomes unnoticed. Our proposal is to adjust speech leakage power below the masking threshold so that the speech distortions are minimized.

2. PROBLEM FORMULATION

Consider an array of N microphones with arbitrary geometry and single speech source $s(t)$ located inside reverberant enclosure. The observation signal at nth microphone is given by:

$$x_n(t) = a_n(t) * s(t) + v_n(t) = y_n(t) + v_n(t), \qquad (1)$$

where $*$ denote a convolution operator, a_n is acoustic impulse response from the source speech signal to the nth microphone and $y_n(t)$, $v_n(t)$ are the clean speech and noise components received at nth microphone.

Work supported by Bialystok University of Technology under the grant S/WI/4/08.

The multichannel systems are often implemented in the frequency-domain using the discrete Fourier transform (DFT). The samples are processed on frame-by-frame basis using analysis window of the length N_f. Let $X_n(\omega)$, $A_n(\omega)$, $S(\omega)$, $Y_n(\omega)$ and $V_n(\omega)$ denote the DFTs of $x_n(t)$, $a_n(t)$, $s(t)$, $y_n(t)$ and $v_n(t)$ respectively (frequency bin indices are omitted for clarity). For sufficiently large N_f, we can approximate the model (1) as follows [3]:

$$\mathbf{x}(\omega) = \mathbf{a}(\omega)S(\omega) + \mathbf{v}(\omega) = \mathbf{y}(\omega) + \mathbf{v}(\omega), \tag{2}$$

where

$$\begin{aligned} \mathbf{x}(\omega) &= [X_1(\omega), X_2(\omega), \ldots, X_N(\omega)]^T, \\ \mathbf{a}(\omega) &= [A_1(\omega), A_2(\omega), \ldots, A_N(\omega)]^T, \\ \mathbf{y}(\omega) &= [Y_1(\omega), Y_2(\omega), \ldots, Y_N(\omega)]^T, \\ \mathbf{v}(\omega) &= [V_1(\omega), V_2(\omega), \ldots, V_N(\omega)]^T. \end{aligned} \tag{3}$$

A correlation matrix for an arbitrary vector $\mathbf{z}(\omega)$ can be defined as follows:

$$\mathbf{R}_{\mathbf{zz}}(\omega) = E\{\mathbf{z}(\omega)\mathbf{z}^H(\omega)\}, \tag{4}$$

where $E\{.\}$ is expectation operator and the superscript ^{H} denotes conjugate transpose. We also assume that the speech and noise processes are wide-sense stationary and uncorrelated, i.e.: $\mathbf{R}_{\mathbf{xx}}(\omega) = \mathbf{R}_{\mathbf{yy}}(\omega) + \mathbf{R}_{\mathbf{vv}}(\omega)$.

Our aim is to estimate source speech signal $S(\omega)$. The most straightforward way is to apply a linear filter $\mathbf{h}(\omega)$ to observation vector $\mathbf{x}(\omega)$ for each frequency bin:

$$\hat{Y}(\omega) = \mathbf{h}^H(\omega)\mathbf{x}(\omega). \tag{5}$$

Above formula can be viewed as the frequency domain implementation of the finite-impulse-response (FIR) filter. The derivation of the optimal filter $\mathbf{h}(\omega)$ depends on some criteria which we will investigate in the next sections.

3. GENERALIZED SIDELOBE CANCELLER

The GSC approach assumes that the filtering for each channel can be performed in two orthogonal subspaces. This is equivalent to the following decomposition:

$$\mathbf{h}(\omega) = \mathbf{w}(\omega) - \mathbf{B}(\omega)\mathbf{g}(\omega) \tag{6}$$

where $\mathbf{w}(\omega)$ is a fixed beamformer filter of size N, $\mathbf{B}(\omega)$ is a blocking matrix of size $N \times (N-1)$ that spans the null space of $\mathbf{a}(\omega)$ and $\mathbf{g}(\omega)$ is a noise cancellation filter of size $N-1$.

The minimum norm solution for vector $\mathbf{w}(\omega)$, which results in distortionless beamformer, is given by:

$$\mathbf{w}_{\mathbf{a}}(\omega) = \mathbf{a}(\omega)/||\mathbf{a}(\omega)||^2. \tag{7}$$

For example, in the case of delay-and-sum beamformer we have:

$$\mathbf{a}(\omega) = [\mathrm{e}^{-j\omega\tau_1}, \mathrm{e}^{-j\omega\tau_2} \ldots, \mathrm{e}^{-j\omega\tau_N}]^T, \tag{8}$$

where $\tau_1, \tau_2, \ldots, \tau_N$ are relative delays between microphones.

The choice of $\mathbf{B}(\omega)$ is not unique and any matrix satisfying the condition $\mathbf{B}^H(\omega)\mathbf{a}(\omega) = 0$, is able to block the speech and create noise reference signal. For example, a proper blocking matrix can be obtained using true channel transfer-function ratios:

$$\mathbf{B}_{\mathbf{a}}(\omega) = \begin{bmatrix} -\frac{A_2^*(\omega)}{A_1^*(\omega)} & -\frac{A_3^*(\omega)}{A_1^*(\omega)} & \cdots & -\frac{A_N^*(\omega)}{A_1^*(\omega)} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{9}$$

The objective of the GSC approach is to find optimal noise cancellation vector $\mathbf{g}(\omega)$. It can be done by solving the following (unconstrained) optimization problem:

$$\min_{\mathbf{g}(\omega)} E\{|\mathbf{w}^H(\omega)\mathbf{v}(\omega) - \mathbf{g}^H(\omega)\mathbf{B}^H(\omega)\mathbf{v}(\omega)|^2\}. \tag{10}$$

This is equivalent to minimizing average residual noise power at the GSC output. An explicit solution for (10) is multichannel Wiener filter [3]:

$$\mathbf{g}_{\mathrm{W}}(\omega) = [\mathbf{B}^H(\omega)\mathbf{R}_{\mathbf{vv}}(\omega)\mathbf{B}(\omega)]^{-1}\mathbf{B}^H(\omega)\mathbf{R}_{\mathbf{v}}(\omega)\mathbf{w}(\omega). \tag{11}$$

Recalling the decomposition (6), the output of the GSC beamformer can be written as follows:

$$\hat{Y}(\omega) = \hat{Y}_{\mathrm{FBF}}(\omega) - \hat{Y}_{\mathrm{NC}}(\omega), \tag{12}$$

where

$$\begin{aligned} \hat{Y}_{\mathrm{FBF}}(\omega) &= \mathbf{w}^H(\omega)\mathbf{x}(\omega), \\ \hat{Y}_{\mathrm{NC}}(\omega) &= \mathbf{g}^H(\omega)\mathbf{B}^H(\omega)\mathbf{x}(\omega). \end{aligned} \tag{13}$$

It is worthwhile to note that computationally efficient, adaptive implementations are preferred [3]. However in our experiments we use non-recursive implementation for simplicity.

4. SPEECH LEAKAGE CONSTRAINED METHOD

A major drawback of the GSC beamformer is a high sensitivity to model uncertainties. For example the delay-and-sum beamformer is reliable in less-reverberant environments. True channel transfer-functions can be roughly estimated using second-order statistics [3], [6] but in general it is a difficult task. Therefore in our approach, we assume a presence of the estimation errors in the model, explicitly. The output of the GSC beamformer can be decomposed as follows:

$$\hat{Y}(\omega) = \hat{S}(\omega) - \hat{S}_{\mathrm{N}}(\omega) + \hat{V}(\omega) - \hat{V}_{\mathrm{N}}(\omega) \tag{14}$$

where

$$\begin{aligned} \hat{S}(\omega) &= \mathbf{w}^H(\omega)\mathbf{a}(\omega)S(\omega), \\ \hat{V}(\omega) &= \mathbf{w}^H(\omega)\mathbf{v}(\omega), \\ \hat{S}_{\mathrm{N}}(\omega) &= \mathbf{g}^H(\omega)\mathbf{B}^H(\omega)\mathbf{a}(\omega)S(\omega), \\ \hat{V}_{\mathrm{N}}(\omega) &= \mathbf{g}^H(\omega)\mathbf{B}^H(\omega)\mathbf{v}(\omega). \end{aligned} \tag{15}$$

are the beamformer speech component, beamformer noise component, speech leakage and noise reference respectively. If $\mathbf{w}(\omega) \neq \mathbf{w_a}(\omega)$, the speech component is not perfectly dereverberated i.e., $\hat{S}(\omega) \neq S(\omega)$. Similarly, if $\mathbf{B}^H(\omega)\mathbf{a}(\omega) \neq 0$, the speech signal leakages to the noise cancellation loop i.e., $\hat{S}_{\mathrm{N}}(\omega) \neq 0$, which usually results in the cancellation of the speech components at the output of the GSC beamformer. It is difficult to improve dereverberation efficiency, however we can minimize the speech leakage effect at expense of some residual noise increase.

Let's define average power of residual noise and speech leakage respectively at the output of the GSC beamformer:

$$\begin{aligned} \epsilon^2_{\mathbf{v}}(\omega) &= E\{|\hat{V}(\omega) - \hat{V}_{\mathrm{N}}(\omega)|^2\}, \\ \epsilon^2_{\mathbf{s}}(\omega) &= E\{|\hat{S}_{\mathrm{N}}(\omega)|^2\}. \end{aligned} \tag{16}$$

Optimization problem for the GSC method can be reformulated as follows:

$$\min_{\mathbf{g}(\omega)} \epsilon^2_{\mathbf{v}}(\omega), \text{ subject to: } \epsilon^2_{\mathbf{s}}(\omega) = \alpha(\omega), \tag{17}$$

where $\alpha(\omega)$ is a some predefined level of the speech leakage power. The complex Lagrange functional is given by:

$$L(\mathbf{g}(\omega), \lambda(\omega)) = \epsilon^2_{\mathbf{v}}(\omega) + \lambda(\omega)[\epsilon^2_{\mathbf{s}}(\omega) - \alpha(\omega)]. \tag{18}$$

Differentiating (18) with respect to $\mathbf{g}(\omega)$ and equating to zero we find the solution:

$$\mathbf{g}_{\mathrm{SLC}}(\omega) = \mathbf{M}(\omega)^{-1}\mathbf{B}^H(\omega)\mathbf{R_{vv}}(\omega)\mathbf{w}(\omega), \tag{19}$$

where

$$\begin{aligned} \mathbf{M}(\omega) &= \mathbf{B}^H(\omega)[\mathbf{R_{vv}}(\omega) + \lambda(\omega)\mathbf{R_{yy}}(\omega)]\mathbf{B}(\omega) \\ &= \mathbf{B}^H(\omega)[(1-\lambda(\omega))\mathbf{R_{vv}}(\omega) + \lambda(\omega)\mathbf{R_{xx}}(\omega)]\mathbf{B}(\omega). \end{aligned}$$

The Lagrange multiplier $\lambda(\omega)$ provides a trade-off between speech leakage and noise reduction. It can be easily verified that for $\lambda(\omega) \to \infty$ speech leakage power is decreased at the expense of increased residual noise. If $\lambda(\omega) = 0$, the conventional GSC method is obtained.

The simplest approach is to set this parameter to empirically chosen fixed value. However an optimal (from the perceptual point of view) solution is to find λ_{opt} such that the speech distortion is inaudible and the residual noise is as low as possible. It can be done by substituting the masking threshold of the clean speech - $\phi_{\mathrm{m}}(\omega)$ for $\alpha(\omega)$ and solving the optimization constraint (17), i.e.:

$$\mathbf{g}^H(\omega)\mathbf{B}^H(\omega)\mathbf{R_{yy}}(\omega)\mathbf{B}(\omega)\mathbf{g}(\omega) = \phi_{\mathrm{m}}(\omega), \tag{20}$$

In this way the speech distortions can be effectively reduced. This situation is also depicted in the Fig. 1. The derivation of an explicit expression for $\lambda(\omega)$ seems to be a difficult task. Theoretically it can be done numerically but we found that for certain cases the solution may not exists or be unstable (i.e., when the masking threshold level is very small). Therefore instead trying to solve (20) explicitly, we propose a suboptimal solution:

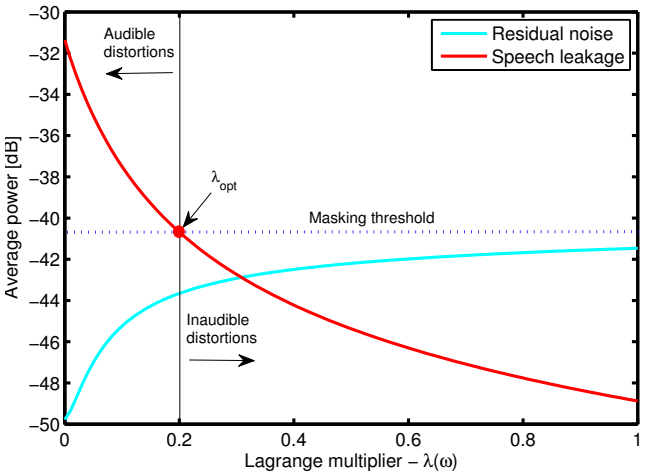

Fig. 1. Speech leakage masking for example speech frame. The curve estimates (16) are obtained using (19).

$$\lambda(\omega) = \lambda_{\mathrm{max}} \mathrm{min}(\mathrm{MNR}(\omega), 1), \tag{21}$$

where λ_{max} is some limiting factor and

$$\mathrm{MNR}(\omega) = \frac{\phi_{\mathrm{m}}(\omega)}{E\{|\hat{V}(\omega)|^2\}} = \frac{\phi_{\mathrm{m}}(\omega)}{\mathbf{w}^H(\omega)\mathbf{R_{vv}}(\omega)\mathbf{w}(\omega)} \tag{22}$$

is the mask to noise ratio. Note that if the noise power level at beamformer output is below the masking threshold ($\mathrm{MNR}(\omega) \geq 1$) the noise is not audible, thus there is no need for noise cancellation and the speech leakage may be minimized as much as possible. Otherwise, if $0 < \mathrm{MNR}(\omega) < 1$, the noise is audible, thus $\lambda(\omega)$ is scaled proportionally to the MNR value, giving a better noise attenuation. Also note that the higher the value of λ_{max} the lower speech distortions. However since the matrix $\mathbf{R_{yy}}(\omega)$ is usually semi-positive definite, the matrix $\mathbf{M}(\omega)$ in (19) may be rank deficient, especially at very high signal-to-noise ratios (SNRs). Therefore the limiting the Lagrange multiplier improves a numerical stability of the inverse in (19). In our experiments λ_{max} was empirically set to 0.25.

Instead of using the MNR one can use a local SNR, but it is known that the SNR estimate is rather erroneous and says nothing about masking effects. In general using the masking threshold is a more robust choice [7]. Most psychoacoustic models compute $\phi_{\mathrm{m}}(\omega)$ by performing some smoothing operations on speech power spectrum. Therefore we estimate the clean speech power spectral density (PSD), first:

$$\phi_{\mathrm{s}}(\omega) \approx E\{|\hat{S}(\omega)|^2\} = \mathbf{w}^H(\omega)\mathbf{R_{yy}}(\omega)\mathbf{w}(\omega). \tag{23}$$

Then we use (23) as an input for Johnston's psychoacoustic model [8]. The correlation matrix of the microphone speech signal is computed as $\mathbf{R_{yy}}(\omega) = \mathbf{R_{xx}}(\omega) - \mathbf{R_{vv}}(\omega)$.

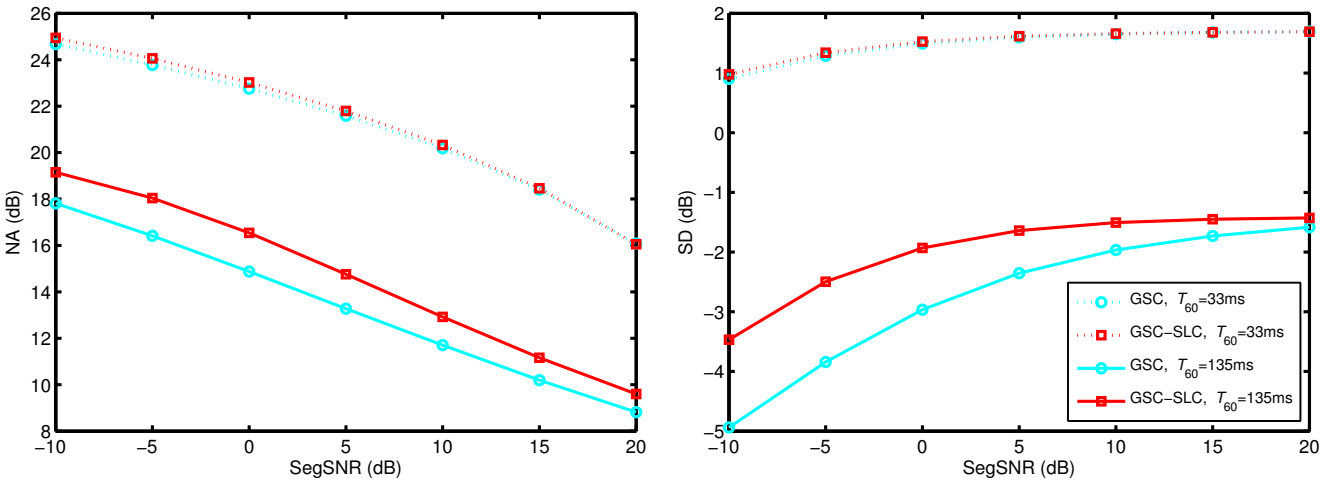

Fig. 2. Objective measures for $N = 8$ microphones: noise attenuation (left), speech distortion (right).

Table 1. Perceptual evaluation using PESQ.

	$T_{60} = 33$ms		$T_{60} = 135$ms	
SegSNR	GSC	GSC-SLC	GSC	GSC-SLC
-10	2.350	2.412	1.797	1.800
-5	2.575	2.648	1.914	1.992
0	2.775	2.847	2.041	2.139
5	2.947	3.006	2.150	2.262
10	3.113	3.165	2.249	2.343
15	3.293	3.356	2.334	2.387
20	3.479	3.566	2.404	2.423

5. EXPERIMENTS

In this section we compare the performance of the conventional GSC beamformer with the proposed speech leakage constrained approach (denoted as GSC-SLC). The methods were implemented in MATLAB using overlap-save procedure. The microphone signals are cut into 50% overlapping frames of size $N_f = 1024$ samples. Once the signals are filtered in the DFT domain they are transformed back to time domain and only last $N_f/2$ samples are saved. In order to determine the system performance under model uncertainties we assumed simple delay-and-sum beamformer (8), thus the steering vector and blocking matrix were computed using (7) and (9), accordingly. To efficiently compute the frequency filters the correlation matrix of the noise signal have to be estimated. However for comparative purposes we put aside this problem and compute $\mathbf{R}_{\mathbf{vv}}(\omega)$ directly from data. In practice any voice activity detector (VAD) can be used to update noise statistics in speech pauses only. Similarly we estimate microphone delays for delay-and-sum beamformer using an exact value of the direction of arrival (DOA) angle.

Two acoustic environments were simulated: the first one with absorptive surfaces ($T_{60} = 33$ms) and the second one with reflective surfaces ($T_{60} = 135$ms). In both cases we assumed the rectangular enclosure with dimensions $6 \times 5 \times 2.8$ (all dimensions and coordinates are in meters). We also considered an uniform linear array of 8 microphones placed on the x-axis with the first microphone at the position $[2.65, 4, 1]$ and spacing 0.05. The speech source signals were sampled at 8kHz and located at the position $[1, 1, 1.8]$. The test material was comprised of 8 sentences, each about 5s long, uttered both by male and female (English spoken) speakers. The white Gaussian noise source was positioned at $[5, 2, 2]$. The microphone signals were obtained by convolving the speech source signal with the room impulse responses and adding to the corresponding noise signal at different SNRs, according to model (1).

In the experiments the SNR based measures were used for an objective performance evaluation. The speech distortion measure (SD) is defined as the segmental signal to noise ratio where the noise is interpreted as a difference between the source signal and enhanced speech. The higher the value of this factor, the better the performance. The amount of noise reduction was measured using noise attenuation factor (NA) defined as the mean ratio between the input noise power and output noise power. Additionally PESQ measure [9] was used for the evaluation of the speech distortion audibility.

The objective measurement results are depicted in Fig. 2. The spectrograms of an example enhancement are presented in Fig. 3. In the case of reverberant environment the proposed method outperforms conventional one providing lower speech distortions. In order to avoid overestimation of noise attenuation factor, it should be measured in speech pauses only, however it is rather difficult to precisely mark this regions. Thus, this factor was estimated also in transients where mean squared error is substantially lower for the speech leakage

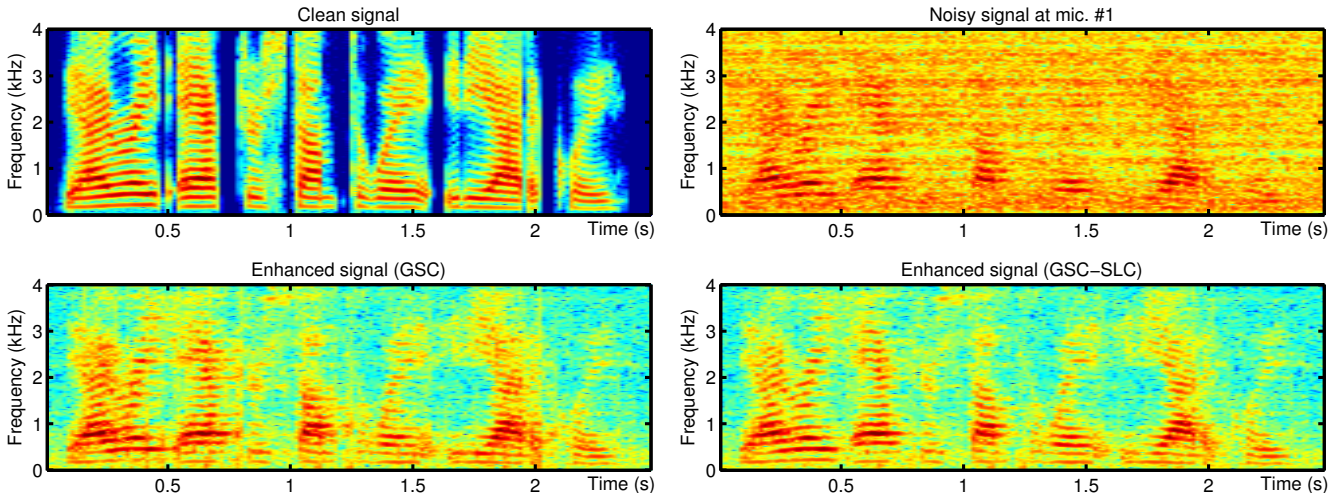

Fig. 3. Spectrograms for reverberant environment ($T_{60} = 135$ms) and noisy speech at SegSNR = 0dB.

constrained method. In fact this measure should be comparable for both methods. On the other hand, it is clear that a residual noise increase is not proportional to the speech distortion decrease. In our experiments this increase is too small to be measured. For non-reverberant environment we reported only minimal improvement. It is not surprising since there is no model uncertainties, thus speech leakage is very low. Thus in this scenario the parameter $\lambda(\omega)$ has no impact on the system performance and the proposed method is equivalent to conventional GSC beamformer. Similar observations can be made for the PESQ scores (see Tab. 1). Although we observe lower performance results for the conventional GSC beamformer for both environment conditions, in the case of reverberant environment (i.e., presence of the system model uncertainties) relative improvement is higher.

6. CONCLUSION

The performance of the conventional GSC beamformer can be improved in the presence system model uncertainties by using auditory properties. We derived a noise cancellation filter which is able to reduce the speech leakage (and speech distortions) at expense of residual noise increase. However as we show this increase is rather small. In addition it is tolerated by auditory system as long as the noise level is placed below masking threshold. The experimental results show that the proposed method outperforms conventional GSC beamformer providing lower speech distortions and comparable residual noise level.

There are some possible improvements of the proposed method, i.e.: a derivation of an explicit formula for optimal Lagrange multiplier, a recursive implementation of the frequency filters or an estimation of the steering vector and blocking matrix using second-order statistics only. These issues will be considered in the future work.

7. REFERENCES

[1] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," in *Proc. IEEE*, Aug 1972, vol. 60, pp. 926–935.

[2] J. Capon, "High resolution frequency-wavenumber spectrum analysis," in *Proc. IEEE*, Aug 1969, vol. 57, pp. 1408–1418.

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

[4] Y. Huang, J. Benesty, and J. Chen, "Analysis and comparison of multichannel noise reduction methods in a common framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 957–968, 2008.

[5] A. Borowicz and A. Petrovsky, "Signal subspace approach for psychoacoustically motivated speech enhancement," *Speech Comm.*, vol. 53, no. 2, pp. 210–219, 2011.

[6] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[7] D. Virette, P. Scalart, and C. Lamblin, "Analysis of background noise reduction techniques for robust speech coding," in *Proc. EUSIPCO*, 2002, vol. 3, pp. 297–300.

[8] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. on Selected Areas in Comm.*, vol. 6, pp. 314–323, February 1988.

[9] ITU-T, "Perceptual evaluation of speech quality (PESQ)," Rec. P.862, ITU, Geneva, 2001.