14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

NOISE ROBUST RELATIVE TRANSFER FUNCTION ESTIMATION

M. Schwab, P. Noll, and T. Sikora

Technical University Berlin, Germany
Communication System Group
Einsteinufer 17, 10557 Berlin (Germany)
{schwab|noll|sikora}@nue.tu-berlin.de

ABSTRACT

Microphone arrays are suitable for a large range of applications. Two important applications are speaker localization and speech enhancement. For both of these the transfer functions from one microphone to the other microphones are needed to form potential algorithms for these applications. In this paper we present a new transfer function estimator optimized for speech sources in a noisy environment. To achieve this, we integrate a new covariance matrix estimation algorithm for the noisy speech as well as for the adaptive and correlated noise signals as received by the microphones. Results indicate that our algorithm outperforms other state-of-the-art algorithms.

1. INTRODUCTION

System identifications is a very useful tool for signal processing tasks. In particular, in speech signal processing system identification is needed for various applications. Two important applications are beamforming algorithms [1] where the system identification is used to form the blocking matrix in adaptive beamformer, and acoustic source localization where system identification is used to determine the time difference of arrival (TDOA) as a prior step to the localization [2],[3].

The traditional method for estimating a transfer system is the cross-correlation method. This method has the disadvantage that it is biased in the presence of noise. Weinstein and Shalvi proposed in [4] an unbiased estimator of the transfer function by assuming non-stationarity of the desired signal whereas the cross-correlation between the additive interfering signals at the sensors has to be stationary. Then they divide the observed time interval into subintervals and analyze in each subinterval the cross power spectral density (PSD). Hence, they get an equation for each subinterval leading to an overdetermined set of linear equations. With a weighted least-squares (WLS) approach, the error variance is minimized. Unless the problem of the biased estimator was solved the error variance of the estimator is increased because of the reduced number of samples used for the estimation of the cross PSD.

Cohen extended the algorithm from Weinstein and Shalvi and adapted the algorithm to noisy speech signals [5]. Therefore, he used a single channel noise reduction algorithm [6]. He reduces the error variance of the cross PSD estimation with a recursive smoothing algorithm. This increases the correlation between the cross PSD estimation at adjacent subintervals and the obtained equations will be correlated. In this paper we introduce a new estimation algorithm for the covariance matrices optimized for noisy speech signals. Our new estimation algorithm derives an estimation of the covariance matrix of the noisy speech signals as well as for the covariance matrix of the additive noise. We incorporate all the informations comprised in these covariance matrices to improve the estimation of an unbiased transfer function.

2. PROBLEM FORMULATION

Consider a two-microphone system with:

$$\begin{aligned} x_1(t) &= h_1(t) * s(t) + n_1(t) \\ x_2(t) &= h_2(t) * s(t) + n_2(t) \end{aligned} \tag{1}$$

where $x_1(t)$ and $x_2(t)$ represent the microphone signals. $s(t)$ is the desired speech signal. $h_1(t)$ and $h_2(t)$ are modeling the impulse response functions from a point source (in this case a speaker) to the microphone 1 and 2, respectively. The additive noise signals in the microphone signals $n_1(t)$ and $n_2(t)$ comprise a correlated part and an uncorrelated part:

$$\begin{aligned} n_1(t) &= h_{n_1}(t) * n(t) + n_{d_1}(t) \\ n_2(t) &= h_{n_2}(t) * n(t) + n_{d_2}(t) \end{aligned} \tag{2}$$

where $n(t)$ is a directed noise source with the transfer functions to the microphones $h_{n_1}(t)$ and $h_{n_2}(t)$. The operation $*$ denotes the convolution. The parts $n_{d_1}(t)$ and $n_{d_1}(t)$

are uncorrelated diffuse noise. The speech source $s(t)$ is assumed to be uncorrelated with the additive noise parts $n_1(t)$ and $n_2(t)$.

2.1. Relative transfer function system identification

In the relative transfer function system identification one microphone signal is used as a reference signal and is subtracted from the filtered second microphone signal. Figure 1 shows such a system where $x_2(t)$ is used as the reference signal.

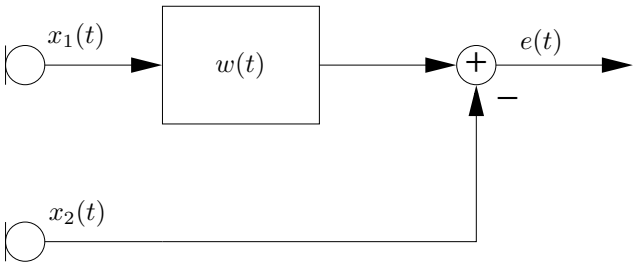

Fig. 1. Reference based system identification

The aim of the system identification (calculation of $w(t)$) is that the error signal

$$e(t) = w(t) * x_1(t) - x_2(t) \tag{3}$$

does not contain any part of the speech source signal $s(t)$. Rewriting equation 3 using equation 1 yields:

$$e(t) = \Big(w(t) * h_1(t) - h_2(t)\Big) * s(t) + w(t) * n_1(t) - n_2(t) \tag{4}$$

To eliminate the speech signal $s(t)$ in $e(t)$ the system identification has to estimate

$$w(t) = h_2(t) * h_1(t)^{-1} \ . \tag{5}$$

Thus, $w(t)$ is the transfer function from the first microphone to the second microphone in the time domain.

Transforming equation 5 into the frequency domain we obtain the relative transfer function (RTF):

$$W(\omega) = \frac{H_2(\omega)}{H_1(\omega)} \ . \tag{6}$$

The simplest algorithm to identify the RTF is the so called cross-correlation method:

$$\hat{W}_{CC}(\omega) = \frac{\phi_{x_1x_2}(\omega)}{\phi_{x_1x_1}(\omega)} \tag{7}$$

in which $\phi_{x_1x_2}(\omega) = E\{X_1^*(\omega)X_2(\omega)\}$ is the cross power spectral density (PSD) of the two microphone signals and $\phi_{x_1x_1}(\omega) = E\{|X_1(\omega)|^2\}$ is the auto power spectral density of the first microphone signal. $X_1^*(\omega)$ is the complex conjugate of $X_1(\omega)$. With the above made assumptions, cross PSD and auto PSD become (for simplicity the parameter ω is omitted in the following):

$$\begin{aligned} \phi_{x_1x_2} &= H_1^* H_2 \phi_{ss} + \phi_{n_1n_2} \\ &= H_1^* H_2 \phi_{ss} + H_{n_1}^* H_{n_2} \phi_{nn} \end{aligned} \tag{8}$$

$$\begin{aligned} \phi_{x_1x_1} &= |H_1|^2 \phi_{ss} + \phi_{n_1n_1} \\ &= |H_1|^2 \phi_{ss} + |H_{n_1}|^2 \phi_{nn} + \phi_{n_{d_1}n_{d_1}} \end{aligned} \tag{9}$$

Equation 8 and 9 show that $\hat{W}_{CC}(\omega)$ is a biased estimator of W in a noisy environment. The cross correlation method is still biased even if only diffuse noise (n_{d_1} and n_{d_2}) is present. For an unbiased estimator of $W(\omega)$ it is necessary to estimate $\phi_{x_1x_2}$, $\phi_{x_1x_1}$, $\phi_{n_1n_2}$, and $\phi_{n_1n_1}$. Then an unbiased estimator is given by:

$$\hat{W}^1_{unbiased} = \frac{\phi_{x_1x_2} - \phi_{n_1n_2}}{\phi_{x_1x_1} - \phi_{n_1n_1}} \tag{10}$$

another unbiased estimation of W is:

$$\hat{W}^2_{unbiased} = \frac{\phi_{x_2x_2} - \phi_{n_2n_2}}{\phi_{x_2x_1} - \phi_{n_2n_1}} \tag{11}$$

where the estimations of $\phi_{x_2x_2}$ and $\phi_{n_2n_2}$ are also used. The final unbiased estimation is then mean of both:

$$\hat{W}_{unbiased} = (f^1 \hat{W}^1_{unbiased} + f^2 \hat{W}^2_{unbiased}) \tag{12}$$

where f^1 and f^2 are weighting functions depending on the SNR. They are defined as follows:

$$f^i(SNR) = \frac{\frac{\phi_{x_ix_i}}{\phi_{n_in_i}}}{\frac{\phi_{x_1x_1}}{\phi_{n_1n_1}} + \frac{\phi_{x_2x_2}}{\phi_{n_2n_2}}} \tag{13}$$

for $i = 1, 2$.

The covariance matrix of the noisy speech signals and covariance matrix of the noise signal contain all desired cross and auto power spectra.

$$\mathbf{C_{x_1x_2}} = \begin{pmatrix} \phi_{x_1x_1} & \phi_{x_1x_2} \\ \phi_{x_2x_1} & \phi_{x_2x_2} \end{pmatrix} \tag{14}$$

$$\mathbf{C_{n_1n_2}} = \begin{pmatrix} \phi_{n_1n_1} & \phi_{n_1n_2} \\ \phi_{n_2n_1} & \phi_{n_2n_2} \end{pmatrix} \tag{15}$$

3. IMPLEMENTATION

In our implementation the system identification is carried out in the frequency domain. Therefore, we use the short-time Fourier Transformation (STFT) where the parameters k and l will denote the frequency and time block index, respectively.

As we have seen in the previous section we need to estimate the covariance matrix of the noisy speech signals and the covariance matrix of the pure noise signals. The estimation is carried out with a recursive smoothing. If $X(k,l)$ and $Y(k,l)$ are the short-time spectrum of the signals $x(n)$ and $y(n)$ then smoothing of the covariance matrix can be expressed as

$$\mathbf{C_{XY}}(\mathbf{k},\mathbf{l}) = \alpha \mathbf{C_{XY}}(\mathbf{k},\mathbf{l}-\mathbf{1}) + (1-\alpha)\mathbf{C^{inst}_{XY}}(k,l) \quad (16)$$

where α is a smoothing factor and $\mathbf{C^{inst}_{XY}}(k,l) = \begin{pmatrix} X^*(k,l) \\ Y^*(k,l) \end{pmatrix} \begin{pmatrix} X(k,l) & Y(k,l) \end{pmatrix}$ is the instantaneous estimation of the covariance matrix.

For the traditional algorithms the smoothing factor is kept constant $\alpha = 0.95$.

In our scenario we deal with speech signals corrupted with background noise. Therefore, we have developed a new optimized covariance matrix estimation for the noisy speech and for the background noise based on a recursive smoothing. In contrast to the previous estimation, where the smoothing factor remains constant, the smoothing factor becomes time and frequency dependent, i. e. $\alpha(k,l)$. Depending on the actual signal-to-noise ratio $SNR(k,l)$ the smoothing factor for the noisy speech can be defined as follows:

$$\alpha_{Opt}(k,l) = \quad (17)$$
$$\begin{cases} 1 & \text{if } SNR(k,l) \leq \text{SNR}_{min} \\ \alpha_{min} & \text{if } SNR(k,l) \geq \text{SNR}_{max} \\ 1-(1-\alpha_{min})\frac{SNR(k,l)-\text{SNR}_{min}}{\text{SNR}_{max}-\text{SNR}_{min}} & \text{otherwise} \end{cases}$$

In [7] an estimator of the a priori SNR was derived which is used here. Finally, the optimized estimation of the noisy speech covariance matrix is given by:

$$\begin{aligned} \mathbf{C^{Opt}_{x_1x_2}}(k,l) &= \alpha_{Opt}(k,l)\mathbf{C^{Opt}_{x_1x_2}}(k,l-1) \quad (18) \\ &\quad +(1-\alpha_{Opt}(k,l))\mathbf{C^{inst}_{x_1x_2}}(k,l) \end{aligned}$$

If $SNR(k,l)$ is low the adaptation of the covariance matrix is stopped because the smoothing factor $\alpha(k,l)$ is set to one. On the other hand, the instantaneous estimation of covariance matrix $\mathbf{C^{inst}_{x_1x_2}}(k,l)$ gets a high weight because $\alpha(k,l)$ is set small. The estimation of the $SNR(k,l)$ is coming from a one channel noise reduction system.

The estimation of the noise covariance matrix is also controlled by the smoothing factor. The speech presence probability $p(k,l)$ defined in [8] controls the update equation for the noise covariance matrix estimation. The algorithm [8] is extended with the optimally smoothed power spectral density estimation according to [9]. The update rule for the noise covariance matrix estimation is proposed by:

$$\begin{aligned} \mathbf{C_{n_1n_2}}(k,l) &= \alpha_N(k,l)\mathbf{C_{n_1n_2}}(k,l-1) \quad (19) \\ &\quad +(1-\alpha_N(k,l))\mathbf{C^{inst}_{x_1x_2}}(k,l) \end{aligned}$$

with

$$\alpha_N(k,l) = p(k,l)(1-\alpha_{min}) + \alpha_{min}\ . \quad (20)$$

Figure 2 illustrates the performance of the used noise tracking algorithm. The noisy signal is a speech signal with additive cockpit noise. The SNR of the noisy signal is 1 dB. The thine line represents the optimally smoothed PSD of the noisy speech [9] and the bold line shows the tracking of the noise PSD.

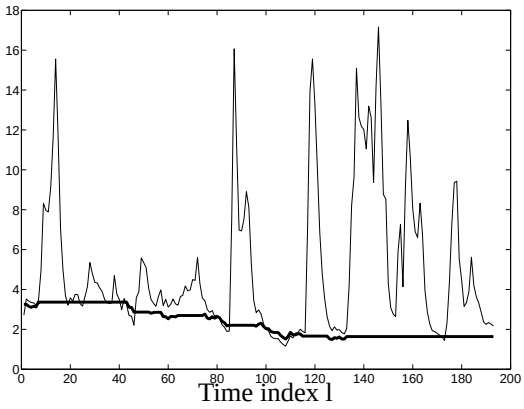

Fig. 2. Non-stationary noise tracking example for the frequency bin $k = 15$. Thin line: optimally smoothed PSD of noisy signal; bold line: estimated noise PSD

4. EXPERIMENTAL SETUP AND RESULTS

For the evaluation of the different system identification algorithms we generated two microphone signals with the following parameters. The different impulse response answers are given by discrete sequences:

$$\begin{aligned} h_1(n) &= [1, 0.4, 0.1, -0.3, 0.2, -0.1, 0, 0] \\ h_2(n) &= [0, 1, 0.5, -0.3, 0, 0.2, 0, 0.1] \\ h_{n_1}(n) &= [0, 1, 0, 0.5, 0.3, 0.1, -0.2] \\ h_{n_2}(n) &= [1, 0, 0.5, 0.1, 0, -0.2, 0] \end{aligned}$$

In figure 3 an example of a reverberated clean speech signal (upper signal waveform) and the noisy speech signal

(lower signal waveform) with a SNR of -2,54 dB are plotted. The additive noise is a stationary white gaussian noise.

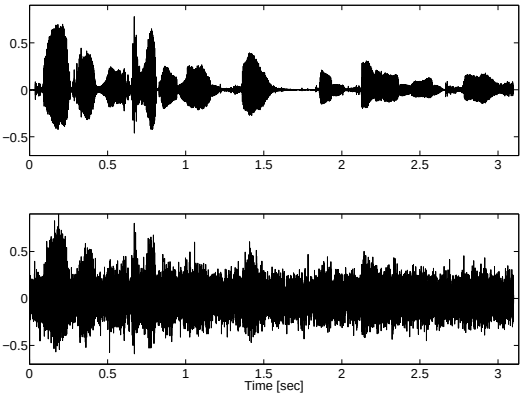

Fig. 3. Signal waveforms, upper plot: reverberated clean speech, lower plot: noisy speech signal

As an objective measure we use the signal blocking factor (SBF) defined in [5] as:

$$\text{SBF} = 10log_{10}\frac{\sum_n (h_1(n) * s(n))^2}{\sum_n e(n)^2} \qquad (21)$$

$\sum_n (h_1(n) * s(n))^2$ is the energy of the speech signal contained in the first microphone signal.

This measure can be pulled up to evaluate the usability of the algorithm for a blocking matrix in an adaptive beamformer.

SBF [dB]	stationary		non-stationary	
global SNR	-0.52 dB	8.97 dB	0.96 dB	8.92 dB
W_{CC}	-0.12	-5.14	-1.71	-7.69
W_{Cohen}	-6.47	-9.63	-8.30	-16.45
$W_{proposed1}$	-10.21	-17.77	-15.11	-22.91
$W_{proposed2}$	-12.54	-18.84	-17.42	-25.01

Table 1. *Comparison of the different system identification algorithms with respect to their SBF performance. W_{CC} is the traditional biased cross correlation method, W_{Cohen} represents the adaptive online algorithm from [5], $W_{proposed1}$ uses the proposed covariance matrix estimation from equation 18, and $W_{proposed2}$ is the implementation of unbiased estimator from equation 12*

Table 1 summarizes the results of the performance evaluation of four different system identification algorithms with respect to their SBF. The tested algorithms are the biased cross correlation method with a constant smoothing factor (W_{CC}), the adaptive online algorithm proposed by Cohen in [5], the cross correlation method with the proposed covariance matrix estimation optimized for noisy speech signals, and the proposed algorithm implementing equation 12. The parameter α_{min} is set to 0.95 and the step size μ in the adaptive online algorithm is chosen to 0.01.

The algorithms are tested with two different noise types, stationary white gaussian and non-stationary cockpit noise. For both noise types two different noise levels are created. The global SNR is the mean SNR between the first and the second microphone signal. In figure 4 the resulting error signals from the different system identification algorithms are visualized. The generated microphone signals are corrupted with the non-stationary noise at a level of 1 dB. The final estimation of the RTF is used for the whole sequence. The adaption process is not considered in the error signals.

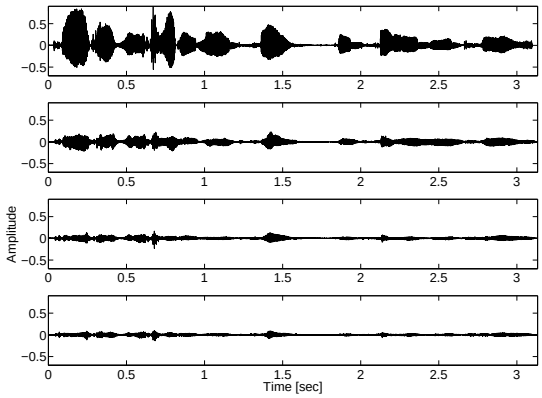

Fig. 4. First plot: reverberated clean signal, second plot: error signal from system identification algorithm after [5], third plot: error signal from with system identification algorith optimized covariance matrix estimation, fourth plot: error signal from unbiased system identification algorithm

5. CONCLUSIONS

In this paper we have introduced a new relative transfer function estimation optimized for noisy speech signals for a two-microphone system. Components from a noise reduction system are used to design an covariance matrix estimation algorithm. It has been shown that this algorithm outperforms other state-of-the-art algorithms. Changes in the transfer function resulting from a speaker movement can be handled since the update rule for the covariance is adaptive. This algorithm is able to cope with competing noise sources. If there is a competing speaker this algorithm will fail because it relies on the assumption to differentiate between noise and speech. In this case a multiple input multiple output (MIMO) system has to be considered.

For future work, we plan to extend the proposed algorithm to a multi-channel system. This will increase the degree of liberty for the identification process. Hence, we expect a better performance in minimizing the final error signal of the system.

6. REFERENCES

[1] S. Gannot, I. Cohen, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *Trans. Signal Processing*, vol. 49, pp. 1614–1626, August 2001.

[2] Yiteng Huang, Jacob Benesty, and Gary W. Elko, "An efficient linear-correction least-squares approach to source localization," in *Applications of Signal Processing to Audio and Acoustics*, October 2001, pp. 67–70.

[3] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *JASA*, vol. 107(1), pp. 384–391, January 2000.

[4] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Transactions Signal Processing*, vol. 44, pp. 2055–2063, August 1996.

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

[6] Israel Cohen, "On speech enhancement under signal presence uncertainty," in *International Conference on Acoustic and Speech Signal Processing*, May 2001, pp. 167–170.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in *IEEE Trans. Acoust., Speech, Signal Processing*, April 1985, vol. ASSP-33, pp. 443–445.

[8] Israel Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[9] Rainer Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, July 2001.