2

Real-Time Dual-Microphone Speech Enhancement

Trabelsi Abdelaziz, Boyer François-Raymond and Savaria Yvon
École Polytechnique de Montréal
Canada

1. Introduction

In various applications such as mobile communications and digital hearing aids, the presence of interfering noise may cause serious deterioration in the perceived quality of speech signals. Thus, there exists considerable interest in developing speech enhancement algorithms that solve the problem of noise reduction in order to make the compensated speech more pleasant to a human listener. The noise reduction problem in single and multiple microphone environments was extensively studied (Benesty et al., 2005; Ephraim. & Malah, 1984). Single microphone speech enhancement approaches often fail to yield satisfactory performance, in particular when the interfering noise statistics are time-varying. In contrast, multiple microphone systems provide superior performance over the single microphone schemes at the expense of a substantial increase of implementation complexity and computational cost.

This chapter addresses the problem of enhancing a speech signal corrupted with additive noise when observations from two microphones are available. It is organized as follows. The next section presents different well-known and state of the art noise reduction methods for speech enhancement. Section 3 surveys the spatial cross-power spectral density (CPSD) based noise reduction approach in the case of a dual-microphone arrangement. Also included in this section, the well known problems associated with the use of the CPSD-based approach. Section 4 describes the single channel noise spectrum estimation algorithm used to cope with the CPSD-based approach shortcomings, and uses this algorithm in conjunction with a soft-decision scheme to come up with the proposed method. We call the proposed method the modified CPSD (MCPSD) based approach. Based on minimum statistics, the noise power spectrum estimator seeks to provide a good tradeoff between the amount of noise reduction and the speech distortion, while attenuating the high energy correlated noise components (i.e., coherent direct path noise), especially in the low frequency range. Section 5 provides objective measures, speech spectrograms and subjective listening test results from experiments comparing the performance of the MCPSD-based method with the cross-spectral subtraction (CSS) based approach, which is a dual-microphone method previously reported in the literature. Finally, Section 6 concludes the chapter.

2. State of the art

There have been several approaches proposed in the literature to deal with the noise reduction problem in speech processing, with varying degrees of success. These approaches

can generally be divided into two main categories. The first category uses a single microphone system and exploits information about the speech and noise signal statistics for enhancement. The most often used single microphone noise reduction approaches are the spectral subtraction method and its variants (O'Shaughnessy, 2000).

The second category of signal processing methods applicable to that situation involves using a microphone array system. These methods take advantage of the spatial discrimination of an array to separate speech from noise. The spatial information was exploited in (Kaneda & Tohyama, 1984) to develop a dual-microphone beamforming algorithm, which considers spatially uncorrelated noise field. This method was extended to an arbitrary number of microphones and combined with an adaptive Wiener filtering in (Zelinski, 1988, 1990) to further improve the output of the beamformer. The authors in (McCowan & Bourlard, 2003) have replaced the spatially uncorrelated noise field assumption by a more accurate model based on an assumed knowledge of the noise field coherence function, and extended the CPSD-based approach to develop a more appropriate postfiltering scheme. However, both methods overestimate the noise power spectral density at the beamformer's output and, thus, they are suboptimal in the Wiener sense (Simmer & Wasiljeff, 1992). In (Lefkimmiatis & Maragos, 2007), the authors have obtained a more accurate estimation of the noise power spectral density at the output of the beamformer proposed in (Simmer & Wasiljeff, 1992) by taking into account the noise reduction performed by the minimum variance distortionless response (MVDR) beamformer.

The generalized sidelobe canceller (GSC) method, initially introduced in (Griffiths & Jim, 1982), was considered for the implementation of adaptive beamformers in various applications. It was found that this method performs well in enhancing the signal-to-noise ratio (SNR) at the beamformer's output without introducing further distortion to the desired signal components (Guerin et al., 2003). However, the achievable noise reduction performance is limited by the amount of incoherent noise. To cope with the spatially incoherent noise components, a GSC based method that incorporates an adaptive Wiener filter in the look direction was proposed in (Fischer & Simmer, 1996). The authors in (Bitzer et al., 1999) have investigated the theoretical noise reduction limits of the GSC. They have shown that this structure performs well in anechoic rooms, but it does not work well in diffuse noise fields. By using a broadband array beamformer partitioned into several harmonically nested linear subarrays, the authors in (Fischer & Kammeyer, 1997) have shown that the resulting noise reduction system performance is nearly independent of the correlation properties of the noise field (i.e., the system is suitable for diffuse as well as for coherent noise field). The GSC array structure was further investigated in (Marro et al., 1998). In (Cohen, 2004), the author proposed to incorporate into the GSC beamformer a multichannel postfilter which is appropriate to work in nonstationary noise environments. To discriminate desired speech transients from interfering transients, he used both the GSC beamformer primary output and the reference noise signals. To get a real-time implementation of the method, the author suggested in an earlier paper (Cohen, 2003a), feeding back to the beamformer the discrimination decisions made by the postfilter.

In the dual-microphone noise reduction context, the authors in (Le Bouquin-Jannès et al., 1997) have proposed to modify both the Wiener and the coherence-magnitude based filters by including a cross-power spectrum estimation to take some correlated noise components into account. In this method, the cross-power spectral density of the two

input signals was averaged during speech pauses and subtracted from the estimated CPSD in the presence of speech. In (Guerin et al., 2003), the authors have suggested an adaptive smoothing parameter estimator to determine the noise CPSD that should be used in the coherence-magnitude based filter. By evaluating the required overestimation for the noise CPSD, the authors showed that the musical noise (resulting from large fluctuations of the smoothing parameter between speech and non-speech periods) could be carefully controlled, especially during speech activity. A simple soft-decision scheme based on minimum statistics to estimate accurately the noise CPSD was proposed in (Zhang & Jia, 2005).

Considering ease of implementation and lower computational cost when compared with approaches requiring microphone arrays with more than two microphones, dual-microphone solutions are yet a promising class of speech enhancement systems due to their simpler array processing, which is expected to lead to lower power consumption, while still maintaining sufficiently good performance, in particular for compact portable applications (i.e., digital hearing aids, and hands-free telephones). The CPSD-based approach (Zelinski, 1988, 1990), the adaptive noise canceller (ANC) approach (Maj et al., 2006), (Berghe & Wooters, 1998), and the CSS-based approach (Guerin et al., 2003; Le Bouquin-Jannès et al., 1997; Zhang & Jia, 2005) are well-known examples. The former lacks robustness in a number of practical noise fields (i.e., coherent noise). The standard ANC method provides high speech distortion in the presence of crosstalk interferences between the two microphones. Formerly reported in the literature, the CSS-based approach provides interesting performance in a variety of noise fields. However, it lacks efficiency in dealing with highly nonstationary noises such as the multitalker babble. This issue will be further discussed later in this chapter.

3. CPSD-based noise reduction approach

This section introduces the signal model and gives a brief review of the CPSD-based approach in the case of a dual-microphone arrangement. Let $s(t)$ be a speech signal of interest, and let the signal vector $n(t) = [n_1(t)\ \ n_2(t)]^T$ denote two-channel noise signals at the output of two spatially separated microphones. The sampled noisy signal $x_m(i)$ observed at the m^{th} microphone can then be modeled as

$$x_m(i) = s(i) + n_m(i), \quad m = 1,\ 2 \tag{1}$$

where i is the sampling time index. The observed noisy signals are segmented into overlapping time frames by applying a window function and they are transformed into the frequency domain using the short-time Fourier transform (STFT). Thus, we have for a given time frame:

$$X(k,l) = S(k,l) + N(k,l) \tag{2a}$$

where k is the frequency bin index, and l is the time index, and where

$$X(k,l) = [X_1(k,l)\ \ X_2(k,l)]^T \tag{2b}$$

$$N(k,l) = [N_1(k,l) \;\; N_2(k,l)]^T \tag{2c}$$

The CPSD-based noise reduction approach is derived from Wiener's theory, which solves the problem of optimal signal estimation in the mean-square error sense. The Wiener filter weights the spectral components of the noisy signal according to the signal-to-noise power spectral density ratio at individual frequencies given by:

$$W(k,l) = \frac{\Phi_{SS}(k,l)}{\Phi_{X_m X_m}(k,l)} \tag{3}$$

where $\Phi_{SS}(k,l)$ and $\Phi_{X_m X_m}(k,l)$ are respectively the power spectral densities (PSDs) of the desired signal and the input signal to the m^{th} microphone.

For the formulation of the CPSD-based noise reduction approach, the following assumptions are made:

1. The noise signals are spatially uncorrelated, $E\{N_1^*(k,l) \cdot N_2(k,l)\} = 0$;
2. The desired signal $S(k,l)$ and the noise signal $N_m(k,l)$ are statistically independent random processes, $E\{S^*(k,l) \cdot N_m(k,l)\} = 0, \;\; m = 1, \; 2$;
3. The noise PSDs are the same on the two microphones.

Under those assumptions, the unknown PSD $\Phi_{SS}(k,l)$ in (3) can be obtained from the estimated spatial CPSD $\Phi_{X_1 X_2}(k,l)$ between microphone noisy signals. To improve the estimation, the estimated PSDs are averaged over the microphone pair, leading to the following transfer function:

$$\hat{W}(k,l) = \frac{\Re\{\hat{\Phi}_{X_1 X_2}(k,l)\}}{(\hat{\Phi}_{X_1 X_1}(k,l) + \hat{\Phi}_{X_2 X_2}(k,l))/2} \tag{4}$$

where $\Re\{\cdot\}$ is the real operator, and " $\hat{}$ " denotes the estimated value. It should be noted that only the real part of the estimated CPSD in the numerator of equation (4) is used, based on the fact that both the auto-power spectral density of the speech signal and the spatial cross-power spectral density of a diffuse noise field are real functions.

There are three well known drawbacks associated with the use of the CPSD-based approach. First, the noise signals on different microphones often hold correlated components, especially in the low frequency range, as is the case in a diffuse noise field (Simmer et al., 1994). Second, such approach usually gives rise to an audible residual noise that has a cosine shaped power spectrum that is not pleasant to a human listener (Le Bouquin-Jannès et al., 1997). Third, applying the derived transfer function to the output signal of a conventional beamformer yields an effective reduction of the remaining noise components but at the expense of an increased noise bias, especially when the number of microphones is too large (Simmer & Wasiljeff, 1992). In the next section, we will focus our attention on estimating and discarding the residual and coherent noise components resulting from the use of the CPSD-based approach in the case of a dual-microphone arrangement. For such system, the overestimation of the noise power spectral density should not be a problem.

4. Dual-microphone speech enhancement system

In this section, we review the basic concepts of the noise power spectrum estimator algorithm on which the MCPSD method presented later, is based. Then, we use a variation of this algorithm in conjunction with a soft-decision scheme to cope with the CPSD-based approach shortcomings.

4.1 Noise power spectrum estimation

For highly nonstationary environments, such as the multitalker babble, the noise spectrum needs to be estimated and updated continuously to allow an effective noise reduction. A variety of methods were recently reported that continuously update the noise spectrum estimate while avoiding the need for explicit speech pause detection. In (Martin, 2001), a method known as the minimum statistics (MS) was proposed for estimating the noise spectrum by tracking the minimum of the noisy speech over a finite window. The author in (Cohen & Berdugo, 2002) suggested a minima controlled recursive algorithm (MCRA) which updates the noise spectrum estimate by tracking the noise-only periods of the noisy speech. These periods were found by comparing the ratio of the noisy speech to the local minimum against a fixed threshold. In the improved MCRA approach (Cohen, 2003b), a different approach was used to track the noise-only periods of the noisy signal based on the estimated speech-presence probability. Because of its ease of use that facilitates affordable (hardware, power and energy wise) real-time implementation, the MS method was considered for estimating the noise power spectrum.

The MS algorithm tracks the minima of a short term power estimate of the noisy signal within a time window of about 1 s. Let $\hat{P}(k,l)$ denote the smoothed spectrum of the squared magnitude of the noisy signal $X(k,l)$, estimated at frequency k and frame l according to the following first-order recursive averaging:

$$\hat{P}(k,l) = \hat{\alpha}(k,l) \cdot \hat{P}(k,l-1) + (1-\hat{\alpha}(k,l)) \cdot |X(k,l)|^2 \tag{5}$$

where $\hat{\alpha}(k,l)$ $(0 < \hat{\alpha}(k,l) < 1)$ is a time and frequency dependent smoothing parameter. The spectral minimum at each time and frequency index is obtained by tracking the minimum of D successive estimates of $\hat{P}(k,l)$, regardless of whether speech is present or not, and is given by the following equation:

$$\hat{P}_{\min}(k,l) = \min(\hat{P}_{\min}(k,l-1), \hat{P}(k,l)) \tag{6}$$

Because the minimum value of a set of random variables is smaller than their average, the noise spectrum estimate is usually biased. Let $B_{\min}(k,l)$ denote the factor by which the minimum is smaller than the mean. This bias compensation factor is determined as a function of the minimum search window length D and the inverse normalized variance $Q_{eq}(k,l)$ of the smoothed spectrum estimate $\hat{P}(k,l)$. The resulting unbiased estimator of the noise spectrum $\hat{\sigma}_n^2(k,l)$ is then given by:

$$\hat{\sigma}_n^2(k,l) = B_{\min}(k,l) \cdot \hat{P}_{\min}(k,l) \tag{7}$$

To make the adaptation of the minimum estimate faster, the search window of D samples is subdivided into U subwindows of V samples ($D = U \cdot V$) and the noise PSD estimate is updated every V subsequent PSD estimates $\hat{P}(k,l)$. In case of a sudden increase in the noise floor, the noise PSD estimate is updated when a local minimum with amplitude in the vicinity of the overall minimum is detected. The minimum estimate, however, lags behind by at most $D + V$ samples when the noise power increases abruptly. It should be noted that the noise power estimator in (Martin, 2001) tends to underestimate the noise power, in particular when frame-wise processing with considerable frame overlap is performed. This underestimation problem is known and further investigation on the adjustment of the bias of the spectral minimum can be found in (Martin, 2006) and (Mauler & Martin, 2006).

4.2 Dual-microphone noise reduction system

Although the CPSD-based method has shown its effectiveness in various practical noise fields, its performance could be increased if the residual and coherent noise components were estimated and discarded from the output spectrum. In the MCPSD-based method, this is done by adding a noise power estimator in conjunction with a soft-decision scheme to achieve a good tradeoff between noise reduction and speech distortion, while still guaranteeing its real-time behavior. Fig. 1 shows an overview of the MCPSD-based system, which is described in details in this section.

We consider the case in which the average of the STFT magnitude spectra of the noisy observations received by the two microphones, $|Y(k,l)| = (|X_1(k,l)| + |X_2(k,l)|)/2$, is multiplied by a spectral gain function $G(k,l)$ for approximating the magnitude spectrum of the sound signal of interest, that is

$$|\hat{S}(k,l)| = G(k,l) \cdot |Y(k,l)| \tag{8}$$

The gain function $G(k,l)$ is obtained by using equation (4), and can be expressed in the following extended form as

$$G(k,l) = \frac{(|X_1(k,l)| \cdot |X_2(k,l)|) \cdot \cos(\Delta\phi(k,l))}{(|X_1(k,l)|^2 + |X_2(k,l)|^2)/2} \tag{9a}$$

where

$$\Delta\varphi(k,l) = \varphi_{X_1}(k,l) - \varphi_{X_2}(k,l) \tag{9b}$$

and where $\varphi_{X_1}(k,l)$ and $\varphi_{X_2}(k,l)$ denote the phase spectra of the STFTs of $X_1(k,l)$ and $X_2(k,l)$ respectively that satisfy the relationship $|\varphi_{X_1}(k,l) - \varphi_{X_2}(k,l)| < \pi/2$. In the implementation of the MCPSD-based approach, any negative values of the gain function $G(k,l)$ are reset to a minimum spectral floor, on the assumption that such frequencies cannot be recovered. Moreover, good results can be obtained when the gain function $G(k,l)$ is squared, which improves the signals selectivity (i.e., those coming from the direct path).

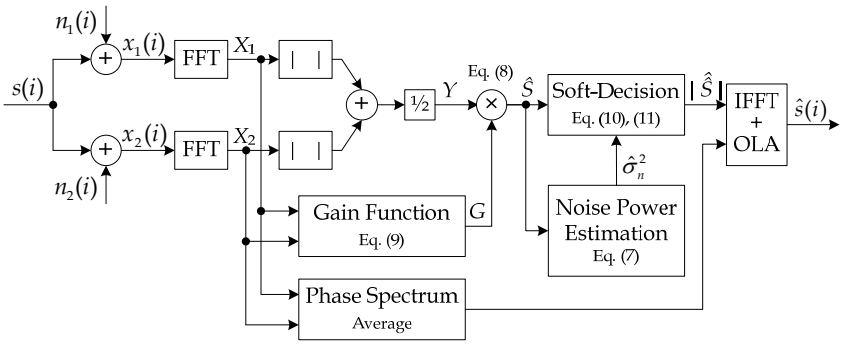


Fig. 1. The proposed dual-microphone noise reduction system for speech enhancement, where "| |" denotes the magnitude spectrum.

To track the residual and coherent noise components that are often present in the estimated spectrum in (8), a variation of the MS algorithm was implemented as follows. In performing the running spectral minima search, the D subsequent noise PSD estimates were divided into two sliding data subwindows of $D/2$ samples. Whenever $D/2$ samples were processed, the minimum of the current subwindow was stored for later use. The sub-band noise power estimate $\hat{\sigma}_n^2(k,l)$ was obtained by picking the minimum value of the current signal PSD estimate and the latest $D/2$ PSD values. The sub-band noise power was updated at each time step. As a result, a fast update of the minimum estimate was achieved in response to a falling noise power. In case of a rising noise power, the update of the minimum estimate was delayed by D samples. For accurate power estimates, the bias correction factor introduced in (Martin, 2001) was scaled by a constant decided empirically. This constant was obtained by performing the MS algorithm on a white noise signal so that the estimated output power had to match exactly that of the driving noise in the mean sense.

To discard the estimated residual and coherent noise components, a soft-decision scheme was implemented. For each frequency bin k and frame index l, the signal to noise ratio was estimated. The signal power was estimated from equation (8) and the noise power was the latest estimated value from equation (7). This ratio, called difference in level (DL), was calculated as follows:

$$DL = 10 \cdot \log_{10}\left(\frac{|\hat{S}(k,l)|^2}{\hat{\sigma}_n^2(k,l)}\right) \tag{10}$$

The estimated DL value was then compared to a fixed threshold Th_s decided empirically. Based on that comparison, a running decision was taken by preserving the sound frequency bins of interest and reducing the noise bins to a minimum spectral floor. That is,

$$|\hat{\tilde{S}}(k,l)| = \begin{cases} |\tilde{S}(k,l)| \cdot \lambda\,, & \text{if } DL < 0 \\ |\tilde{S}(k,l)| \cdot \left(\left(\frac{DL}{Th_s}\right)^2 \cdot (1-\lambda) + \lambda\right), & \text{if } DL < Th_s \\ |\tilde{S}(k,l)|\,, & \text{otherwise.} \end{cases} \tag{11a}$$

where

$$|\tilde{S}(k,l)| = \sqrt{|\hat{S}(k,l)|^2 - \hat{\sigma}_n^2(k,l)} \tag{11b}$$

and where λ was chosen such that $20 \cdot \log_{10}(\lambda) \cong -40$ dB . The argument of the square-root function in equation (11b) was restricted to positive values in order to guarantee real-valued results. When the estimated DL value is lower than the statistical threshold, the quadratic function "$(DL/Th_s)^2 \cdot (1-\lambda) + \lambda$" allows the estimated spectrum to be smoothed during noise reduction. It should be noted that the so called DL has to take positive values during speech activity and negative values during speech pause periods.

Finally, the estimated magnitude spectrum in (11) was combined with the average of the phase spectra of the two received signals prior to estimate the time signal of interest. In addition to the 6 dB reduction in phase noise, the time waveform resulting from such combination provided a better match of the sound signal of interest coming from the direct path. After an inverse DFT of the enhanced spectrum, the resultant time waveform was half-overlapped and added to adjacent processed segments to produce an approximation of the sound signal of interest (see Fig. 1).

5. Performance evaluation and results

This section presents the performance evaluation of the MCPSD-based method, as well as the results of experiments comparing this method with the CSS-based approach. In all the experiments, the analysis frame length was set to 1024 data samples (23 ms at 44.1 kHz sampling rate) with 50% overlap. The analysis and synthesis windows thus had a perfect reconstruction property (i.e., Hann-window). The sliding window length of D subsequent PSD estimates was set to 100 samples. The threshold Th_s was fixed to 5 dB. The recordings were made using a Presonus Firepod recording interface and two Shure KSM137 cardioid microphones placed approximately 20 cm apart. The experimental environment of the MCPSD is depicted in Fig. 2. The room with dimensions of 5.5 x 3.5 x 3 m enclosed a speech source situated at a distance of 0.5 m directly in front (0 degrees azimuth) of the input microphones, and a masking source of noise located at a distance of 0.5 m from the speech source.

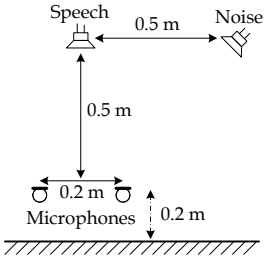


Fig. 2. Overhead view of the experimental environment.

Designed to be equally intelligible in noise, five sentences taken from the Hearing in Noise Test (HINT) database (Nilsson et al., 1994) were recorded at a sampling frequency of 44.1 kHz. They are

1. Sentence 1 (male talker): "Flowers grow in the garden".
2. Sentence 2 (female talker): "She looked in her mirror".
3. Sentence 3 (male talker): "The shop closes for lunch".
4. Sentence 4 (female talker): "The police helped the driver".
5. Sentence 5 (male talker): "A boy ran down the path".

Four different noise types, namely white Gaussian noise, helicopter rotor noise, impulsive noise and multitalker babble noise, were recorded at the same sampling rate and used throughout the experiments. The noise was scaled in power level and added acoustically to the above sentences with a varying SNR. A global SNR estimation of the input data was used. It was computed by averaging power over the whole length of the two observed signals with:

$$\mathrm{SNR} = 10 \cdot \log_{10} \left(\frac{\sum_{m=1}^{2} \sum_{i=1}^{I} s^2(i)}{\sum_{m=1}^{2} \sum_{i=1}^{I} (x_m(i) - s(i))^2} \right) \tag{12}$$

where I is the number of data samples of the signal observed at the m^{th} microphone. Throughout the experiments, the average of the two clean signals $s(i) = (s_1(i) + s_2(i))/2$ was used as the clean speech signal. Objective measures, speech spectrograms and subjective listening tests were used to demonstrate the performance improvement achieved with the MCPSD-based method over the CSS-based approach.

5.1 Objective measures

The Itakura-Saito (IS) distance (Itakura, 1975) and the log spectral distortion (LSD) (Mittal & Phamdo, 2000) were chosen to measure the differences between the clean and the test spectra. The IS distance has a correlation of 0.59 with subjective quality measures (Quakenbush et al., 1988). A typical range for the IS distance is 0–10, where lower values indicate better speech quality. The LSD provides reasonable degree of correlation with subjective results. A range of 0–15 dB was considered for the selected LSD, where the minimum value of LSD corresponds to the best speech quality. In addition to the IS and LSD measures, a frame-based segmental SNR was used which takes into consideration both speech distortion and noise reduction. In order to compute these measures, an utterance of the sentence 1 was processed through the two methods (i.e., the MCPSD and CSS). The input SNR was varied from –8 dB to 8 dB in 4 dB steps.

Values of the IS distance measure for various noise types and different input SNRs are presented in Tables 1 and 2 for signals processed by the different methods. Results in this table were obtained by averaging the IS distance values over the length of sentence 1. The results in this table indicate that the CSS-based approach yielded more speech distortion than that produced with the MCPSD-based method, particularly in helicopter and impulsive noise environments. Fig. 3 illustrates the comparative results in terms of LSD measures between both methods for various noise types and different input SNRs. From these figures, it can be observed that, whereas the two methods showed comparable improvement in the case of impulsive noise, the estimated LSD values provided by the MCPSD-based method

were the lowest in all noise conditions. In terms of segmental SNR, the MCPSD-based method provided a performance improvement of about 2 dB on average, over the CSS-based approach. The largest improvement was achieved in the case of multitalker babble noise, while for impulsive noise this improvement was decreased. This is shown in Fig. 4.

SNR (dB)	White Noise			Helicopter Noise		
	CSS	MCPSD	Noisy	CSS	MCPSD	Noisy
–8	1.88	0.62	3.29	2.81	1.92	3.28
–4	1.4	0.43	2.82	2.18	1.29	2.62
0	0.78	0.3	2.23	1.72	0.95	2.18
4	0.51	0.24	1.64	1.28	0.71	1.7
8	0.34	0.25	1.18	0.87	0.47	1.24

Table 1. Comparative performance in terms of mean Itakura-Saito distance measure for white and helicopter noises and different input SNRs.

SNR (dB)	Impulsive Noise			Babble Noise		
	CSS	MCPSD	Noisy	CSS	MCPSD	Noisy
–8	2.71	2.03	3.23	2.38	1.26	3.1
–4	2.21	1.67	2.65	1.7	0.85	2.62
0	1.7	1.21	2.06	1.28	0.59	2.12
4	1.34	0.93	1.56	0.92	0.46	1.73
8	0.99	0.69	1.09	0.67	0.32	1.27

Table 2. Comparative performance in terms of mean Itakura-Saito distance measure for impulsive and babble noises and different input SNRs.

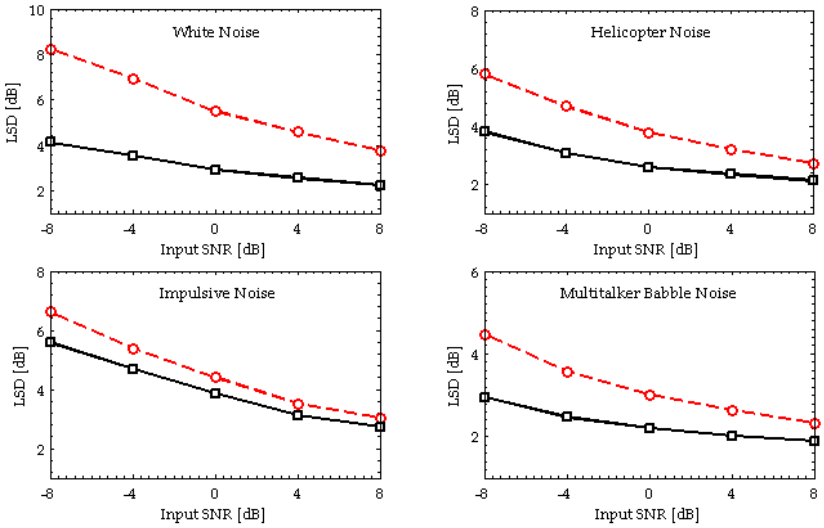


Fig. 3. Log spectral distortion measure for various noise types and levels, obtained using (○) CSS approach, and (□) the MCPSD-based method.

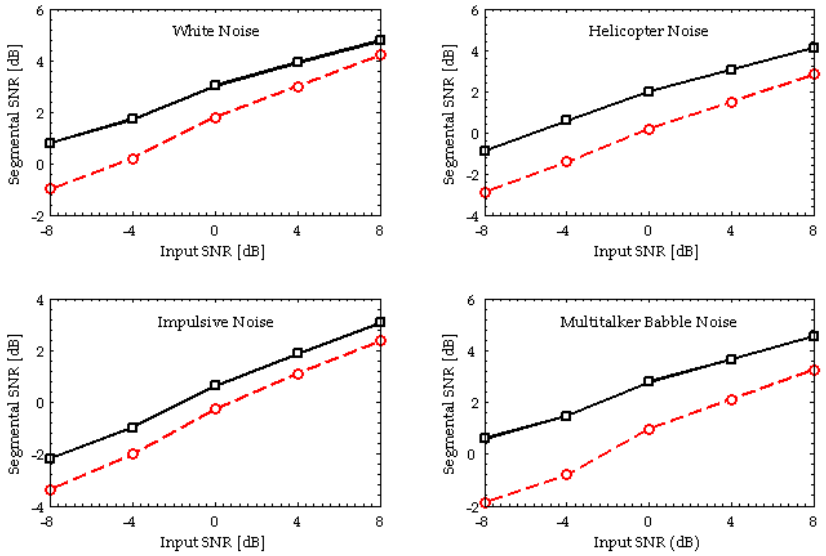


Fig. 4. Segmental SNR improvement for various noise types and levels, obtained using (○) CSS approach and (□) the MCPSD-based method.

5.2 Speech spectrograms

Objective measures alone do not provide an adequate evaluation of system performance. Speech spectrograms constitute a well-suited tool for analyzing the time-frequency behavior of any speech enhancement system. All the speech spectrograms presented in this section (Figs. 5–8) use sentence 1 corrupted with different background noises at SNR = 0 dB.

In the case of white Gaussian noise (Fig. 5), whereas the MCPSD-based method and the CSS-based approach provided sufficient amount of noise reduction, the spectrum of the former preserved better the desired speech components. In the case of helicopter rotor noise (Fig. 6), large residual noise components were observed in the spectrograms of the signals processed by the CSS-based approach. Unlike this method, the spectrogram of the signal processed by the MCPSD-based method indicated that the noise between the speech periods was noticeably reduced, while the shape of the speech periods was nearly unchanged. In the case of impulsive noise (Fig. 7), it can be observed that the CSS-based approach was less effective for this type of noise. In contrast, the spectrogram of the signal processed by the MCPSD-based method shows that the impulsive noise was moderately reduced in both the speech and noise periods. In the case of multitalker babble noise (Fig. 8), it can be seen that the CSS-based approach provided limited noise reduction, particularly in the noise only periods. By contrast, a good noise reduction was achieved by the MCPSD-based method on the entire spectrum.

We can conclude that, while the CSS-based approach afforded limited noise reduction, especially for highly nonstationary noise such as multitalker babble, the MCPSD-based method can deal efficiently with both stationary and transient noises with less spectral distortion even in severe noisy environments.

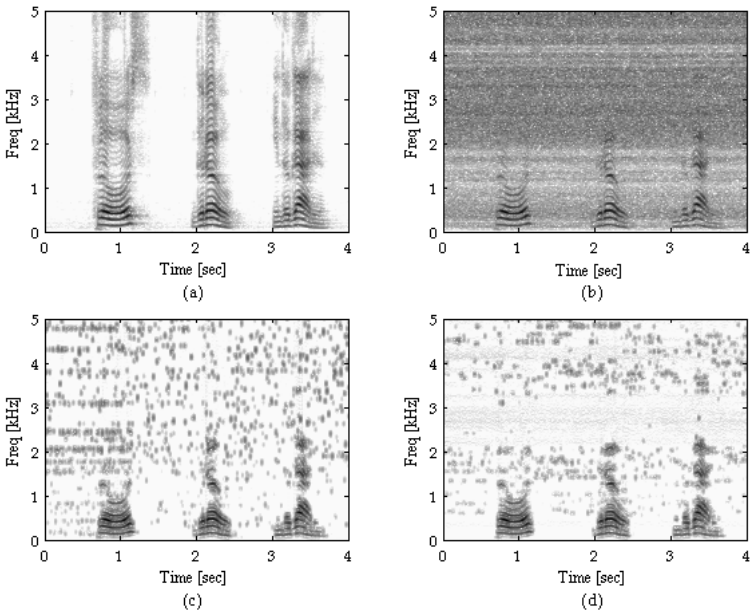


Fig. 5. Speech spectrograms obtained with white Gaussian noise added at SNR=0 dB. (a) Clean speech (b) Noisy signal (c) CSS output (d) MCPSD output.

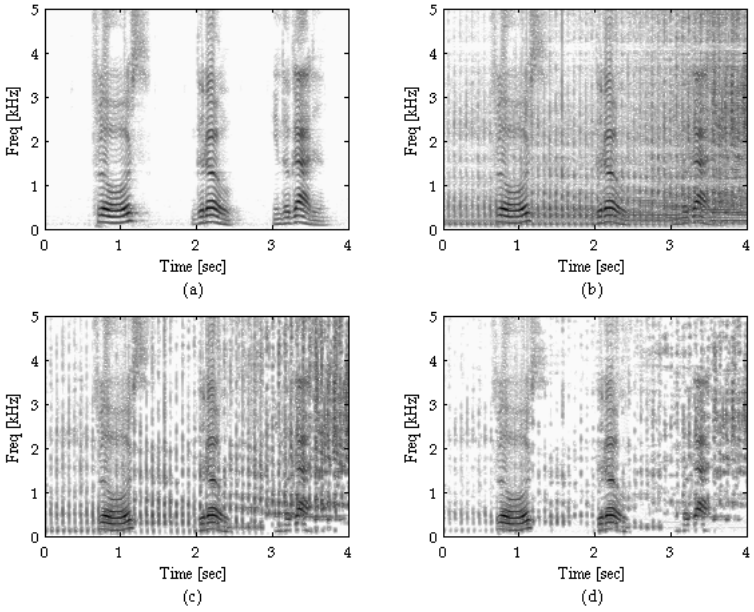


Fig. 6. Speech spectrograms obtained with helicopter rotor noise added at SNR=0 dB. (a) Clean speech (b) Noisy signal (c) CSS output (d) MCPSD output.

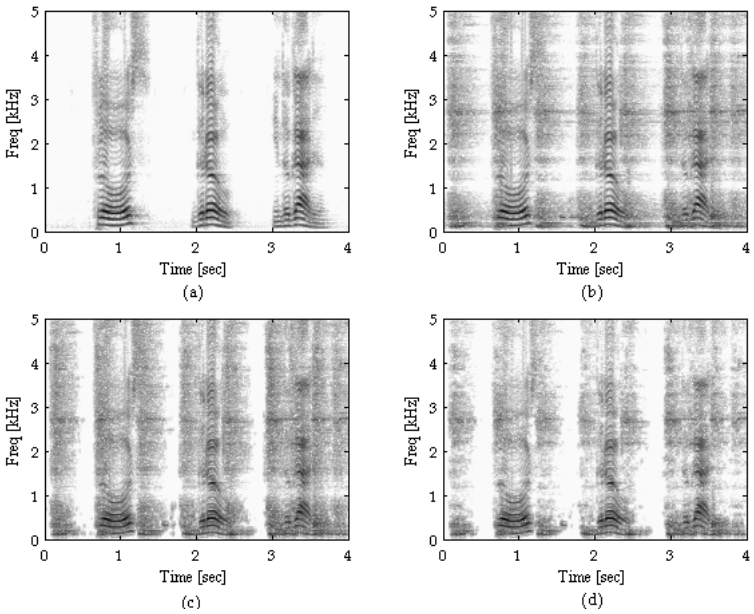


Fig. 7. Speech spectrograms obtained with impulsive noise added at SNR=0 dB. (a) Clean speech (b) Noisy signal (c) CSS output (d) MCPSD output.

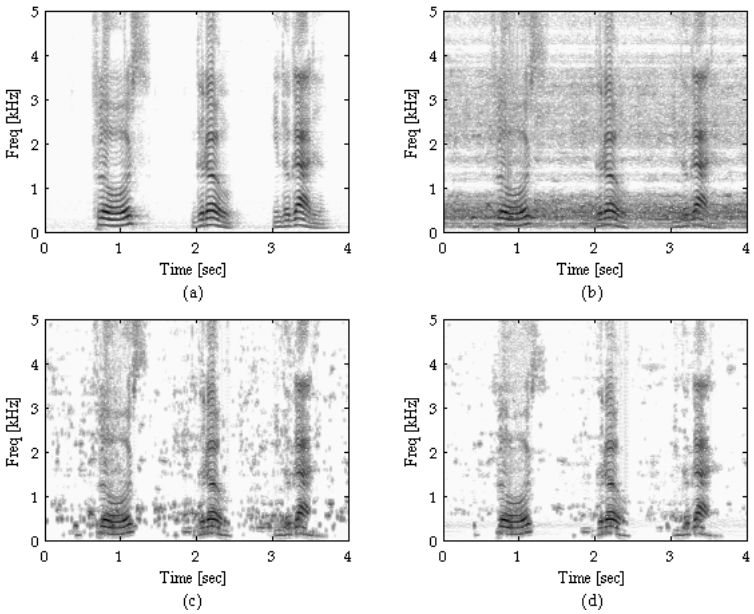


Fig. 8. Speech spectrograms obtained with multitalker babble noise added at SNR=0 dB. (a) Clean speech (b) Noisy signal (c) CSS output (d) MCPSD output.

5.3 Subjective listening tests

In order to validate the objective performance evaluation, subjective listening tests were conducted with the MCPSD and the CSS based approaches. The different noise types considered in this study were added to utterances of the five sentences listed before with SNRs of –5, 0, and 5 dB. The test signals were recorded on a portable computer, and headphones were used during the experiments. The seven-grade comparison category rating (CCR) was used (ITU-T, Recommendation P.800, 1996). The two methods were scored by a panel of twelve subjects asked to rate every sequence of two test signals between –3 and 3. A negative score was given whenever the former test signal sounded more pleasant and natural to the listener than the latter. Zero was selected if there was no difference between the two test signals. For each subject, the following procedure was applied: 1) each sequence of two test signals was played with brief pauses in between tracks and repeated twice in a random order; 2) the listener was then asked if he wished to hear the current sequence once more or skip to the next. This led to 60 scores for each test session which took about 25 minutes per subject. The results, averaged over the 12 listeners' scores and the 5 test sentences, are shown in Fig. 9. For the considered background noises, CCRs ranging from 0.33 to 1.27 were achieved over the alternative approach. The maximum improvement of CCR was obtained in the case of helicopter noise (1.1) and multitalker babble noise (1.27), while the worst score was achieved for additive white noise (0.33). The reason behind the roughly similar performance of the two methods in the case of white noise can be understood by recognizing that the minimum statistics noise PSD estimator performs better in the presence of stationary noise as opposed to nonstationary noise.

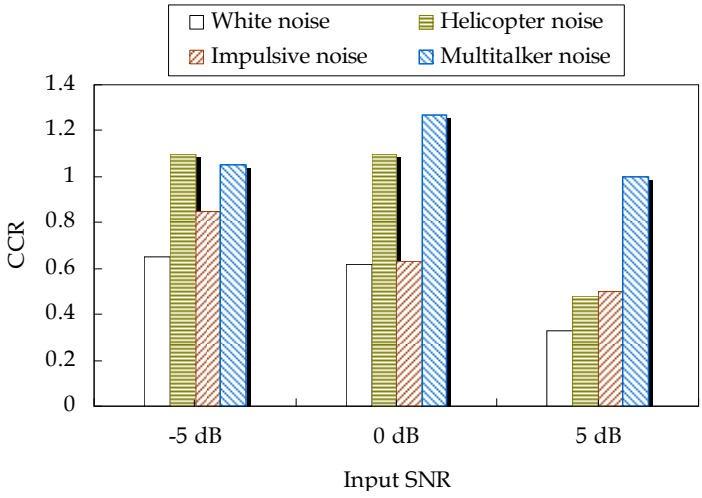


Fig. 9. CCR improvement against CSS for various noise types and different SNRs.

6. Conclusion

Given two received signals corrupted by additive noise, adding a noise power spectrum estimator after the CPSD-based noise reduction system, can substantially reduce the residual and coherent noise components that would otherwise be present at the output spectrum. The

added noise power estimator seeks to provide a good tradeoff between the amount of noise reduction and the speech distortion, while attenuating the high energy correlated noise components, especially in the low frequency ranges. The performance evaluation of the modified CPSD-based method, formerly named MCPSD in this chapter, was carried out over the CSS-based approach, a dual-microphone method previously reported in the literature. Objective evaluation results show that a performance improvement in terms of segmental SNR of about 2 dB on average can be achieved by the MCPSD-based method over the CSS-based approach. The best noise reduction was obtained in the case of multitalker babble noise, while the improvement was lower for impulsive noise. Subjective listening tests performed on a limited data set revealed that CCRs ranging from 0.33 to 1.27 can be achieved over the CSS-based approach. The maximum improvement of CCR was obtained in the case of helicopter and multitalker babble noises, while the worst score was achieved when white noise was added. A fruitful direction of further research would therefore be to extend the MCPSD-based method to multiple microphones as well as to investigate the benefits of such extension on the overall system performance.

7. References

Benesty, J. et al. (2005). Speech Enhancement, Springer, ISBN 978-3540240396, New York, USA.

Berghe, J.V. & Wooters, J. (1998). An adaptive noise canceller for hearing aids using two nearby microphones. *Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3621–3626.

Bitzer, J. et al. (1999). Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. *24th IEEE International Conference on Acoustics, Speech & Signal Processing*, vol. 5, pp. 2965–2968, Phoenix, USA, March 1999.

Cohen, I. & Berdugo, B. (2002). Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Transaction on Signal & Audio Processing*, vol. 9, no. 1, pp. 12–15.

Cohen, I. et al. (2003a). An integrated real-time beamforming and postfiltering system for nonstationary noise environments. *EURASIP Journal on Applied Signal Processing*, pp. 1064–1073.

Cohen, I. (2003b). Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transaction on Speech & Audio Processing*, vol. 11, no. 5, pp. 466–475.

Cohen, I. (2004). Multichannel post-filtering in nonstationary noise environments. *IEEE Transaction on Signal Processing*, vol. 52, no. 5, pp. 1149–1160.

Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transaction on Audio, Speech & Signal Processing*, vol. 32, no. 6, pp. 1109-1121.

Fischer, S. & Simmer, K.U. (1996). Beamforming microphone arrays for speech acquisition in noisy environments. *Speech Communication*, vol. 20, no. 3–4, pp. 215–227.

Fischer, S. & Kammeyer, K.D. (1997). Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environments. *22th IEEE International Conference on Acoustics, Speech & Signal Processing*, vol. 1, pp. 359–362, Munich, Germany, April 1997.

Griffiths, L.J. & Jim, C.W. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transaction on Antennas & Propagation*, vol. 30, no. 1, pp. 27–34.

Guerin, A. et al. (2003). A two-sensor noise reduction system: applications for hands-free car kit. *EURASIP Journal on Applied Signal Processing*, pp. 1125–1134.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transaction on Audio Speech & Signal Processing*, vol. 23, pp. 67-72.

ITU-T, Recommendation P.800 (1996). Methods for subjective determination of transmission quality. *International Telecommunication Union Radiocommunication Assembly*.

Kaneda, Y. & Tohyama, M. (1984). Noise suppression signal processing using 2-point received signal. *Electronics and Communications in Japan*, vol. 67–A, pp. 19–28.

Le Bouquin-Jannès, R. et al. (1997). Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator. *IEEE Transaction on Speech & Audio Processing*, vol. 5, pp. 484–487.

Lefkimmiatis, S. & Maragos, P. (2007). A generalized estimation approach for linear and nonlinear microphone array post-filters. *Speech Communication*, vol. 49, pp. 657–666.

Maj, J.B. et al. (2006). Comparison of adaptive noise reduction algorithms in dual microphone hearing aids. *Speech Communication*, vol. 48, no. 8, pp. 957–970.

Marro, C. et al. (1998). Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transaction on Speech & Audio Processing*, vol. 6, no. 3, pp. 240–259.

Martin, R. (2001). Noise power spectral estimation based on optimal smoothing and minimum statistics. *IEEE Transaction on Signal & Audio Processing*, vol. 9, pp. 504–512.

Martin, R. (2006). Bias compensation methods for minimum statistics noise power spectral density estimation. *Signal Processing*, vol. 86, no. 6, pp. 1215–1229.

Mauler, D. & Martin, R. (2006). Noise power spectral density estimation on highly correlated data. *10th International Workshop on Acoustic, Echo & Noise Control*, Paris, France, September 2006.

McCowan, I.A. & Bourlard, H. (2003). Microphone array post-filter based on noise field coherence. *IEEE Transaction on Speech & Audio Processing*, vol. 11, no. 6, pp. 709–716.

Mittal, U. & Phamdo, N. (2000). Signal/Noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Transaction on Speech & Audio Processing*, vol. 8, no. 2, pp. 159-167.

Nilsson, M. et al. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085-1099.

O'Shaughnessy, D. (2000). *Speech Communications, Human and Machine*, IEEE Press, ISBN 0-7803-3449-3, New York, USA.

Quakenbush, S. et al. (1988). Objective Measures of Speech Quality. Englewood Cliffs, Prentice-Hall, ISBN/ISSN 0136290566, 9780136290568.

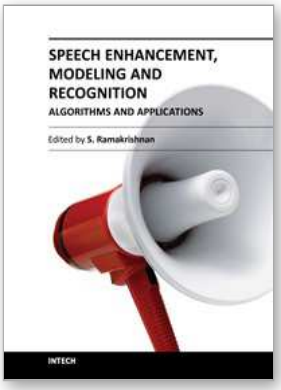
Simmer, K.U. & Wasiljeff, A. (1992). Adaptive microphone arrays for noise suppression in the frequency domain. *Second COST229 Workshop on Adaptive Algorithms in Communications*, pp. 185–194, Bordeaux, France, October 1992.

Simmer, K.U. et al. (1994). Suppression of coherent and incoherent noise using a microphone array. *Annales des Télécommunications*, vol. 49, pp. 439–446.

Zelinski, R. (1988). A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. *13th IEEE International Conference on Acoustics, Speech & Signal Processing*, vol. 5, pp. 2578–2581, NY, USA, April 1988.

Zelinski, R. (1990). Noise reduction based on microphone array with LMS adaptive post-filtering. *Electronic Letters*, vol. 26, no. 24, pp. 2036–2581.

Zhang, X. & Jia, Y. (2005). A soft decision based noise cross power spectral density estimation for two-microphone speech enhancement systems. *IEEE International Conference on Acoustics, Speech & Signal Processing*, vol. 1, pp. I/813–16, Philadelphia, USA, March 2005.



Speech Enhancement, Modeling and Recognition- Algorithms and Applications

Edited by Dr. S Ramakrishnan

ISBN 978-953-51-0291-5
Hard cover, 138 pages
Publisher InTech
Published online 14, March, 2012
Published in print edition March, 2012

This book on Speech Processing consists of seven chapters written by eminent researchers from Italy, Canada, India, Tunisia, Finland and The Netherlands. The chapters covers important fields in speech processing such as speech enhancement, noise cancellation, multi resolution spectral analysis, voice conversion, speech recognition and emotion recognition from speech. The chapters contain both survey and original research materials in addition to applications. This book will be useful to graduate students, researchers and practicing engineers working in speech processing.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Trabelsi Abdelaziz, Boyer François-Raymond and Savaria Yvon (2012). Real-Time Dual-Microphone Speech Enhancement, Speech Enhancement, Modeling and Recognition- Algorithms and Applications, Dr. S Ramakrishnan (Ed.), ISBN: 978-953-51-0291-5, InTech, Available from: http://www.intechopen.com/books/speech-enhancement-modeling-and-recognition-algorithms-and-applications/real-time-dual-microphone-speech-enhancement

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821