MICROPHONE ARRAY POST-FILTERING FOR NON-STATIONARY NOISE SUPPRESSION

Israel Cohen and Baruch Berdugo

Lamar Signal Processing Ltd., P.O.Box 573, Yokneam Ilit 20692, Israel
icohen@lamar.co.il; http://www.AndreaElectronics.com

ABSTRACT

Microphone array post-filtering allows additional reduction of noise components at a beamformer output. Existing techniques are either restricted to classical delay-and-sum beamformers, or are based on single-channel speech enhancement algorithms that are inefficient at attenuating highly non-stationary noise components. In this paper, we introduce a microphone array post-filtering approach, applicable to adaptive beamformer, that differentiates non-stationary noise components from speech components. The ratio between the transient power at the beamformer primary output and the transient power at the reference noise signals is used for indicating whether such a transient is desired or interfering. Based on a Gaussian statistical model and combined with an appropriate spectral enhancement technique, a significantly reduced level of non-stationary noise is achieved without further distorting speech components. Experimental results demonstrate the effectiveness of the proposed method.

1. INTRODUCTION

Microphone array systems are often used for high quality hands-free communication in reverberant and noisy environments [1]. Compared to single microphone systems, a substantial gain in performance is obtainable due to the spatial filtering capability to suppress interfering signals coming from undesired directions. Post-filtering at a beamformer output, based on the Wiener approach, allows further reduction of incoherent noise components [2]–[6]. However, existing post-filtering techniques are either restricted to classical delay-and-sum beamformers, or are based on single-channel speech enhancement algorithms [7]. A single-channel post-filtering approach lacks the ability to attenuate highly non-stationary noise components, since such components are indistinguishable from the speech components.

Recently, we introduced a noise estimation approach, namely *Minima Controlled Recursive Averaging* (MCRA) [8, 9], that is particularly advantageous under low input signal-to-noise ratio (SNR) and non-stationary noise conditions. The noise estimate is obtained by averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability in subbands. The speech presence probability is based on a Gaussian statistical model and controlled by the minima values of a smoothed periodogram of the noisy speech. We have shown that compared to competitive methods, the MCRA noise estimate responses more quickly to noise variations and obtains significantly lower estimation error. When integrated into a speech enhancement system, it yields higher speech quality and a lower level of musical residual noise.

In this paper, we extend the MCRA approach to microphone array post-filtering. By exploiting the relation between the beamformer primary output and the reference noise signals, we make a distinction between non-stationary noise components and speech components. The speech is assumed to be strongest at the primary output, while a noise component is presumably strongest at one of the reference signals. Hence, the ratio between the transient power at the primary output and the transient power at the reference signals is used for indicating whether such a transient is desired or interfering. The speech presence probability determines the rate of recursive averaging for obtaining a noise estimate. When speech is present, the recursive averaging is slow, thus preventing the noise estimate from increasing as a result of speech activity. As the probability of speech presence decreases, the recursive averaging rate increases, facilitating a faster update of the noise estimate. Combined with an appropriate spectral enhancement technique, we achieve a significantly reduced level of non-stationary noise without further distorting speech components.

The paper is organized as follows. In Section 2, we review the MCRA noise estimation approach. The microphone array post-filtering is introduced in Section 3. Experimental results, which validate the effectiveness of the proposed method, are presented in Section 4.

2. SINGLE MICROPHONE NOISE SPECTRUM ESTIMATION

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, and let $y(n)$ represent the noisy observed signal. In the short-term Fourier domain we have $Y(k,\ell) = X(k,\ell) + D(k,\ell)$, where k designates the frequency bin index, and ℓ the frame index. The MCRA approach for noise spectrum estimation [8, 9] is to recursively average past spectral power values of the noisy measurement, using a smoothing parameter that is controlled by the minima values of a smoothed periodogram. The recursive averaging is given by

$$\hat{\lambda}_d(k,\ell+1) = \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) + \beta\cdot[1-\tilde{\alpha}_d(k,\ell)]|Y(k,\ell)|^2 \quad (1)$$

where $\hat{\lambda}_d(k,\ell)$ is an estimate for the noise spectrum $E\left\{|D(k,\ell)|^2\right\}$, $\tilde{\alpha}_d(k,\ell)$ is a time-varying frequency-

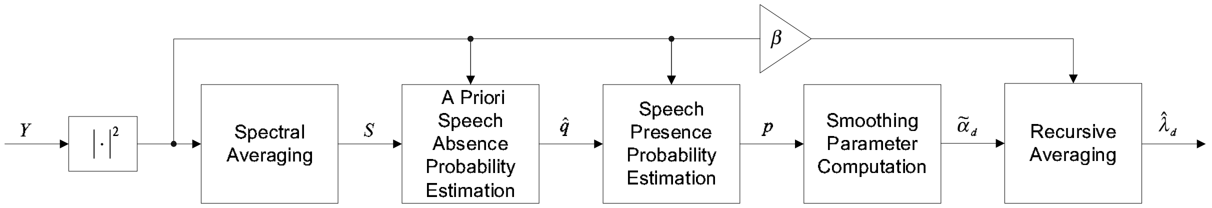

Y
$|\cdot|^2$
Spectral Averaging
S
A Priori Speech Absence Probability Estimation
$\hat{q}$
Speech Presence Probability Estimation
p
Smoothing Parameter Computation
$\tilde{\alpha}_d$
Recursive Averaging
$\hat{\lambda}_d$
β

Fig. 1. Block diagram of the MCRA noise spectrum estimation.

dependent smoothing parameter, and β is a factor that compensates the bias when speech is absent [9]. The smoothing parameter is determined by the speech presence probability, $p(k,\ell)$, and a constant α_d $(0 < \alpha_d < 1)$ that represents its minimal value:

$$\tilde{\alpha}_d(k,\ell) \triangleq \alpha_d + (1-\alpha_d)p(k,\ell)\,. \tag{2}$$

When speech is present, $\tilde{\alpha}_d$ is close to one, thus preventing the noise estimate from increasing as a result of speech activity. As the probability of speech presence decreases, the smoothing parameter gets smaller, facilitating a faster update of the noise estimate.

Assuming a Gaussian statistical model [10], the speech presence probability is given by

$$p(k,\ell) = \left\{1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell))\exp(-v(k,\ell))\right\}^{-1} \tag{3}$$

where $\gamma(k,\ell) \triangleq |Y(k,\ell)|^2/\lambda_d(k,\ell)$ and $\xi(k,\ell) \triangleq E\left\{|X(k,\ell)|^2\right\}/\lambda_d(k,\ell)$ are respectively the *a posteriori* and *a priori* SNRs, $q(k,\ell)$ is the *a priori* probability for speech absence, and $v \triangleq \gamma\xi/(1+\xi)$.

The *a priori* SNR is estimated by [8]

$$\hat{\xi}(k,\ell) = \alpha G_{H_1}^2(k,\ell-1)\gamma(k,\ell-1) + (1-\alpha)\max\{\gamma(k,\ell)-1, 0\} \tag{4}$$

where α is a weighting factor that controls the trade-off between noise reduction and speech distortion, and

$$G_{H_1}(k,\ell) \triangleq \frac{\xi(k,\ell)}{1+\xi(k,\ell)} exp\left(\frac{1}{2}\int_{v(k,\ell)}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{5}$$

is the spectral gain function of the *Log-Spectral Amplitude* (LSA) estimator when speech is surely present[1] [11]. The *a priori* speech absence probability is estimated by [9]

$$\hat{q}(k,\ell) = \begin{cases} 1, & \text{if } \gamma_{\min}(k,\ell) \le 1 \\ & \quad\text{and } \zeta(k,\ell) < \zeta_0 \\ \frac{\gamma_0 - \gamma_{\min}(k,\ell)}{\gamma_0 - 1}, & \text{if } 1 < \gamma_{\min}(k,\ell) < \gamma_0 \\ & \quad\text{and } \zeta(k,\ell) < \zeta_0 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $\gamma_{min}(k,\ell) \triangleq |Y(k,\ell)|^2/(B_{min}S_{min}(k,\ell))$, $\zeta(k,\ell) \triangleq S(k,\ell)/(B_{min}S_{min}(k,\ell))$, γ_0 and ζ_0 are constants satisfying a certain significance level,

$$S(k,\ell) = \alpha_s S(k,\ell-1) + (1-\alpha_s)\sum_{i=-w}^{w} b(i)\,|Y(k-i,\ell)|^2$$

denotes a smoothed spectrogram of the noisy signal, α_s $(0 < \alpha_s < 1)$ is a smoothing parameter, b is a normalized window function of length $2w+1$ (*i.e.*, $\sum_{i=-w}^{w} b(i) = 1$),

$$S_{min}(k,\ell) \triangleq \min\left\{S(k,\ell') \mid \ell - D + 1 \le \ell' \le \ell\right\}$$

is a running minimum of $S(k,\ell)$ using a finite length window of D frames, and B_{min} is a constant independent of the noise power spectrum such that

$$E\{S_{min}(k,\ell) \mid \xi(k,\ell) = 0\} = B_{min}^{-1}\cdot\lambda_d(k,\ell)\,.$$

A block diagram of the MCRA noise spectrum estimation is shown in Fig. 1. Typical values of parameters used in the implementation of the MCRA algorithm, for a sampling rate of 16 kHz, are: $\alpha_d = 0.85$; $\beta = 1.47$; $\alpha = 0.92$; $\alpha_s = 0.9$; $\gamma_0 = 3$; $\zeta_0 = 1.67$; $w = 1$; $D = 120$; $B_{min} = 1.66$. In this case, the short-term Fourier transform is implemented with Hamming windows of 512 samples length (32 ms) and 128 samples frame update step.

The main advantage of estimating the *a priori* speech absence probability by combining conditions on both $\gamma_{\min}(k,\ell)$ and $\zeta(k,\ell)$ is the exclusion of speech components from the averaging process, hence the prevention of an increase in the estimated noise during weak speech activity. Speech components are generally determined by the condition on $\zeta(k,\ell)$. Some speech components are so weak that $\zeta(k,\ell)$ is smaller than ζ_0. In that case, the condition on $\gamma_{\min}(k,\ell)$ becomes useful. The remaining speech components can hardly affect the noise estimator, since their power is relatively low compared to that of the noise.

3. MICROPHONE ARRAY POST-FILTERING

In this section, we extend the MCRA approach to microphone array post-filtering. While a conventional post-filter is effective for handling pseudo-stationary noise at the output of a beamformer, highly non-stationary noise components are generally not attenuated since they are not differentiated from the speech components. Our objective is to exploit the relation between the beamformer primary output and the reference noise signals in order to make a distinction between non-stationary noise and non-stationary speech.

[1] Notice that under speech presence uncertainty, $\hat{\xi}(k,\ell)$ differs from the "decision-directed" estimator of Ephraim and Malah [10]. Its advantage, particularly for weak speech components and low input SNR, is discussed in [8].

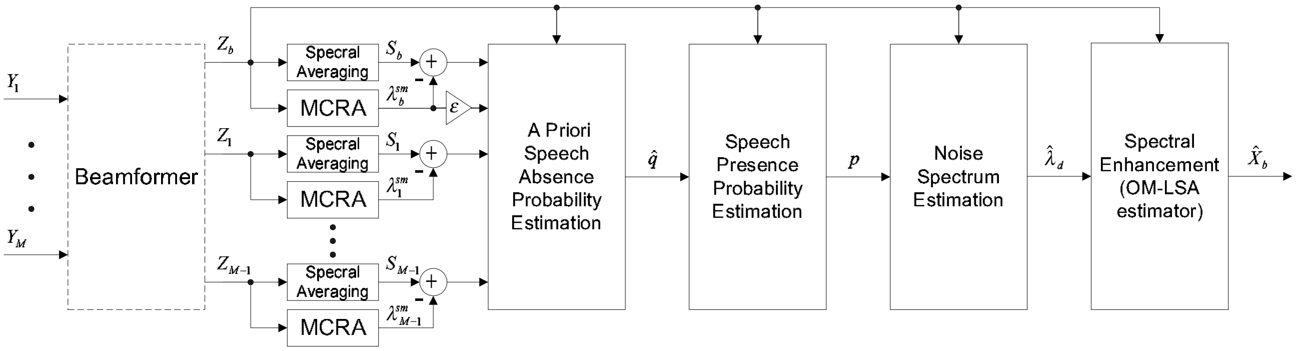


Fig. 2. Block diagram of the microphone array post-filtering.

Fig. 2 shows a block diagram of the proposed microphone array post-filtering. The beamformer primary output Z_b and the reference noise signals $\{Z_i\}_{i=1}^{M-1}$ (where M is the number of microphones in the array) are obtained by adaptively aiming the beamformer at the desired speech source and noise sources, respectively [7]. Each signal,

$$\begin{aligned} Z_b(k,\ell) &= X_b(k,\ell) + D_{bs}(k,\ell) + D_{bt}(k,\ell) \qquad (7)\\ Z_i(k,\ell) &= X_i(k,\ell) + D_{is}(k,\ell) + D_{it}(k,\ell),\ i = 1,\ldots,M-1 \end{aligned}$$

comprises three components. The first is a non-stationary component due to the desired speech signal. The other two are stationary and transient noise components. The speech is presumably strongest at the primary output. On the other hand, a noise component is strongest at one of the reference signals. Hence, the ratio between the transient power at the primary output and the transient power at the reference signals may be used for indicating whether such a transient is desired or interfering.

Let $S_b(k,\ell)$ and $\{S_i(k,\ell)\}_{i=1}^{M-1}$ denote smoothed spectrograms of the beamformer output signals, and let $\lambda_b^{sm}(k,\ell)$ and $\{\lambda_i^{sm}(k,\ell)\}_{i=1}^{M-1}$ represent the respective estimates of the noise spectrum by the MCRA method. Then, the transient beam-to-reference ratio (TBRR) is defined by

$$\psi(k,\ell) = \frac{\max\{S_b(k,\ell) - \lambda_b^{sm}(k,\ell), 0\}}{\max\left\{\{S_i(k,\ell) - \lambda_i^{sm}(k,\ell)\}_{i=1}^{M-1}, \varepsilon\lambda_b^{sm}(k,\ell)\right\}} \qquad (8)$$

where ε (typically $\varepsilon = 0.01$) prevents the denominator from decreasing to zero in the absence of a transient power at the reference signals. The *a priori* speech absence probability is estimated by

$$\hat{q}(k,\ell) = \begin{cases} 1, & \text{if } \gamma_{\min}(k,\ell) \leq 1 \text{ or } \psi(k,\ell) < \psi_{low} \\ \max\left\{\frac{\gamma_0 - \gamma_{\min}(k,\ell)}{\gamma_0 - 1}, \frac{\psi_{high} - \psi(k,\ell)}{\psi_{high} - \psi_{low}}, 0\right\}, & \text{otherwise,} \end{cases} \qquad (9)$$

where ψ_{low} and ψ_{high} are constants that represent the uncertainty in $\psi(k,\ell)$ during weak speech activity (typically $\psi_{low} = 1$, $\psi_{high} = 3$). Substituting $\hat{q}(k,\ell)$ into (3) and computing a smoothing parameter by (2), we obtain the following estimate for the noise spectrum at the beamformer primary output:

$$\hat{\lambda}_d(k,\ell+1) = \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) + \beta[1 - \tilde{\alpha}_d(k,\ell)]|Z_b(k,\ell)|^2 . \qquad (10)$$

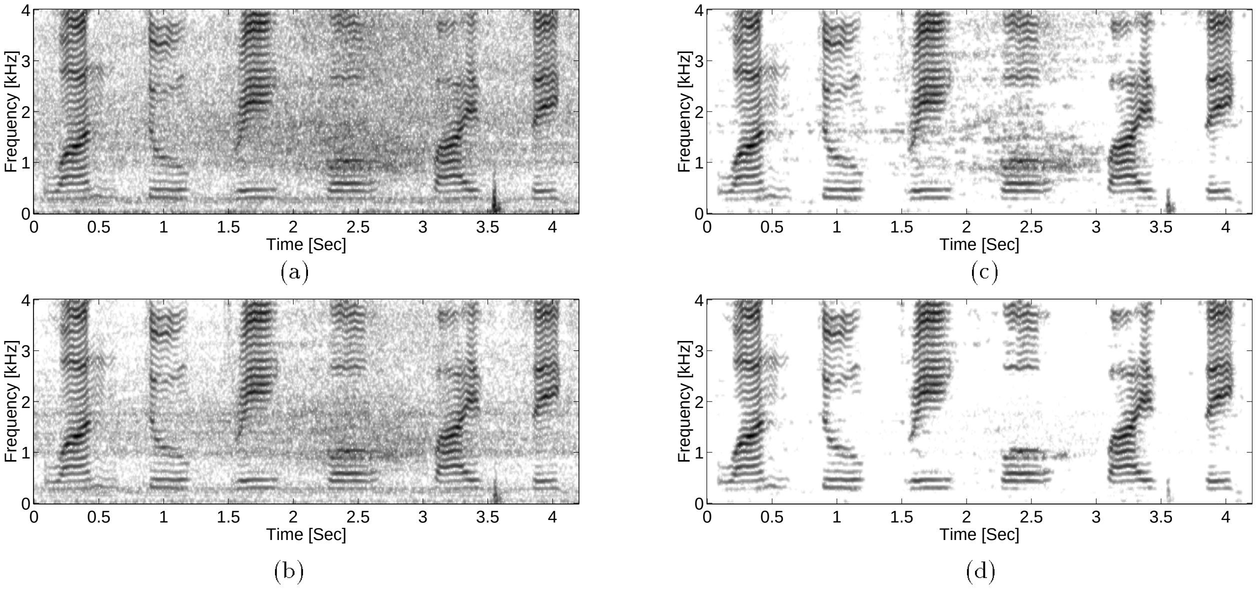
This estimate takes into account the transient, as well as stationary, noise components. Fed into an appropriate spectral enhancement algorithm (*e.g.*, the *Optimally-Modified Log-Spectral Amplitude* (OM-LSA) estimation technique [8]), we achieve a robust estimate for the desired speech component, $\hat{X}_b(k,\ell)$. In particular, improved noise suppression capability is obtained, even under adverse conditions (low SNR, incoherent or diffuse noise fields, highly non-stationary babble, factory or car noise, *etc.*), while retaining weak speech components and avoiding the musical residual noise phenomena.

4. EXPERIMENTAL RESULTS AND DISCUSSION

To validate the usefulness of the proposed post-filtering approach under non-stationary noise conditions, we compare its performance to a single-channel post-filtering in a car environment. Specifically, the speech at the beamformer primary output is enhanced using the OM-LSA estimator, while the noise spectrum is obtained either by the method described in the previous section or by the single-channel MCRA technique.

A linear array, consisting of four microphones with 5 cm spacing, is mounted in a car on the visor. The clean speech signals are recorded at a sampling rate of 16 kHz in the absence of background noise (standing car, silent environment). The noise signals are separately recorded when the windows are slightly open (about 5 cm), and the car speed is about 60 km/h. The input microphone signals are generated by mixing the speech and noise signals at various SNR levels.

Fig. 3 demonstrates the capability of the microphone-array post-filtering to handle abrupt changes in the noise spectrum. Trace of the increase in segmental SNR, gained by the microphone array post-filtering over the single-channel post-filtering is depicted in Fig. 4. The statistics of background noise varies substantially due to a passing car. Additionally, a short burst of low frequency noise follows at about 3.6 sec. The improvement in performance over the single-channel post-filtering is obtained when the noise spectrum fluctuates. In some instances the improvement in SNR surpasses as much as 8 dB. Clearly, a single-channel post-filter is inefficient at attenuating highly non-stationary noise components, since it lacks the ability to differentiate such components from the speech components. On



(a) (c) (b) (d)

Fig. 3. Speech spectrograms. (a) Noisy signal at a single microphone (car noise, SegSNR = 0 dB); (b) Beamformer output (SegSNR = 4.2 dB); (c) Beamformer output enhanced by a single-channel post-filtering (SegSNR = 10.1 dB); (d) Beamformer output enhanced by a microphone array post-filtering (SegSNR = 12.8 dB).

the other hand, the microphone-array post-filter achieves a significantly reduced level of background noise, whether stationary or not, without further distorting speech components. This is verified by subjective informal listening tests.

5. ACKNOWLEDGEMENT

The authors thank Yehuda Mittelman for useful discussions.

6. REFERENCES

[1] M. S. Brandstein and D. B. Ward (eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, 2001.

[2] C. Marro, Y. Mahieux and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. SAP*, vol. 6, no. 3, pp. 240–259, May 1998.

[3] S. Fischer and K.-D. Kammeyer, "Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environments," *Proc. ICASSP*, Munich, Germany, April 1997, pp. 359–362.

[4] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," *Proc. ICASSP*, Munich, Germany, April 1997, pp. 21–24.

[5] J. Bitzer, K. U. Simmer and K.-D. Kammeyer, "Multimicrophone noise reduction by post-filter and superdirective beamformer," *Proc. IWAENC*, Pocono Manor, Pennsylvania, September 1999, pp. 100–103.

[6] I. A. McCowan, C. Marro and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1723–1726.

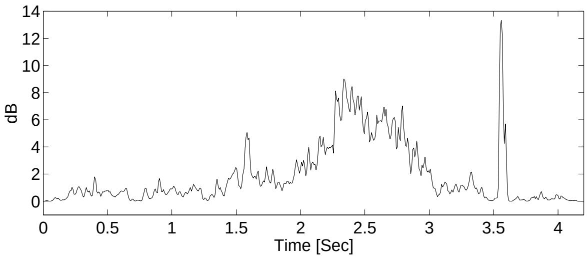


Fig. 4. Trace of the increase in segmental SNR, gained by the microphone array post-filtering over a single-channel post-filtering.

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

[8] I. Cohen and B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, October 2001.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," submitted to *IEEE Trans. SAP*.

[10] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. ASSP*, vol. ASSP-32, pp. 1109–1121, 1984.

[11] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. ASSP*, vol. ASSP-33, pp. 443–445, 1985.