Speech Enhancement Based on a Microphone Array and Log-Spectral Amplitude Estimation

Israel Cohen
Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
icohen@ee.technion.ac.il

Baruch Berdugo
Lamar Signal Processing Ltd.
Andrea Electronics Corp. - Israel
P.O.Box 573, Yokneam Ilit 20692, Israel
bberdugo@lamar.co.il

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. In cases of spatially incoherent noise fields, beamforming alone does not provide sufficient noise reduction, and postfiltering is normally required (see [2, 3] and references therein). Existing microphone array systems are based on beamforming and multi-channel Wiener postfiltering. However, a Wiener filter minimizes the mean-square error (MSE) distortion of the signal estimate, which is essentially not the optimal criterion for enhancing noisy speech. A more appropriate distortion measure for speech enhancement systems is based on the MSE of the spectral, or log-spectral, amplitude [4, 5]. Furthermore, abrupt transient interferences are not attenuated, since the postfilter is unable to track and adapt to fast changes in the noise statistics. Single-channel postfiltering techniques also lack the ability to attenuate transient noise, since transients are generally not differentiated from the desired speech components.

In this paper, we present a multi-microphone speech enhancement approach for minimizing the log-spectral amplitude (LSA) distortion in non-stationary noise environments. An adaptive beamformer with a generalized sidelobe canceller structure is applied to the noisy observed signals. In addition to the beamformer primary output, it provides reference noise signals by projecting the input signals onto the noise-only subsapce. Presumably, a desired signal component is stronger at the beamformer output than at any reference noise signal, and a noise component is strongest at one of the reference signals. Hence, the ratio between the transient power at beamformer output and the transient power at the reference signals indicates whether such a transient is desired or interfering. Based on a Gaussian statistical model [4], and an appropriate decision-directed *a priori* SNR estimate [6], we derive an estimator for the signal presence probability. This estimator controls the rate of recursive averaging in obtaining a noise spectrum estimate by the *Minima Controlled Recursive Averaging* (MCRA) approach [7]. Subsequently, spectral enhancement of the beamformer output is achieved by applying an optimal gain function, which minimizes the MSE of the log-spectra.

2. Problem Formulation

Let $x(t)$ denote a desired speech signal, and let the observed signals at the output of M microphones be given by

$$z_i(t) = a_i(t) * x(t) + n_{is}(t) + n_{it}(t)\,, \quad i = 1, \ldots, M \tag{1}$$

where $a_i(t)$ is the impulse response of the acoustic path between the desired speaker and the i-th microphone, $*$ denotes convolution, n_{is} represents pseudo-stationary noise, and n_{it} represents undesired transient components. An adaptive beamformer (specifically, the *transfer-function generalized sidelobe canceller* (TF-GSC) [8]) is applied to the noisy observed signals. The beamformer comprises a fixed beamformer, which is steered to the look-direction,a blocking matrix, which generates the reference noise signals by projecting the input signals onto the noise-only subspace, and a multi-channel adaptive noise canceller, which reduces the stationary noise that leaks through the sidelobes of the fixed beamformer. We assume that the noise canceller is adapted only to the stationary noise, and not modified during transient interferences. Furthermore, we expect that some desired speech components may leak through the blocking matrix due to steering error.

Using the short-time Fourier transform (STFT), the beamformer primary output and the reference noise signals can be written as

$$\begin{aligned} Y(k,\ell) &= X_1(k,\ell) + D_{1s}(k,\ell) + D_{1t}(k,\ell) \\ U_i(k,\ell) &= X_i(k,\ell) + D_{is}(k,\ell) + D_{it}(k,\ell),\ i = 2, \ldots, M \end{aligned} \tag{2}$$

where the first term in each signal is a non-stationary component due to the desired speech signal. The other two terms are stationary and transient noise components. Our objective is to find an estimator $\hat{X}_1(k,\ell)$ for the desired beamformer output speech, which minimizes the distortion measure

$$E\left\{ \left(\log|X_1(k,\ell)| - \log|\hat{X}_1(k,\ell)| \right)^2 \,|\, Y(k,\ell), U_2(k,\ell), \cdots, U_M(k,\ell \right\}. \tag{3}$$

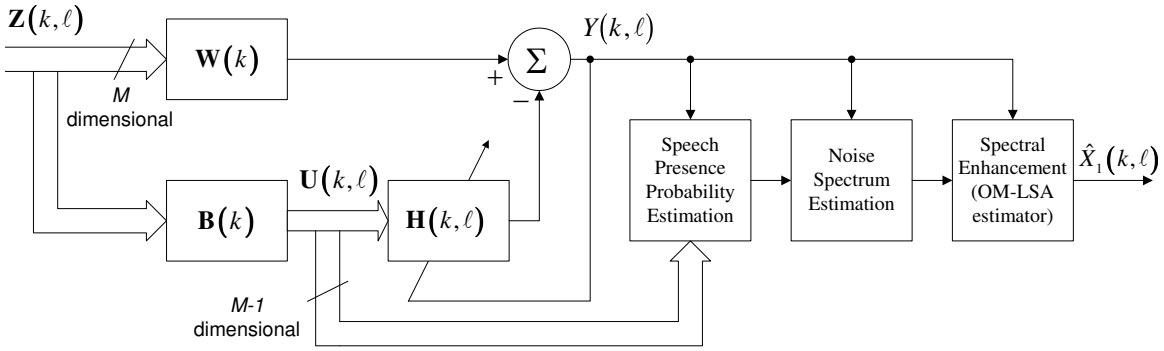

Fig. 1. Block diagram of the multi-microphone log-spectral amplitude estimation.

3. Multi-Microphone Log-Spectral Amplitude Estimation

A block diagram of the proposed multi-microphone log-spectral amplitude estimation scheme is shown in Fig. 1. Transient components are detected at the beamformer output, and an estimate for the signal presence probability is produced based on a Gaussian statistical model. A noise spectrum estimate is obtained by recursively averaging past spectral power values, using a smoothing factor that is adjusted by the speech presence probability. Subsequently, spectral enhancement of the beamformer output is achieved by applying the *optimally-modified log-spectral amplitude* (OM-LSA) gain function [6], which minimizes the MSE of the log-spectra.

Let $\mathcal{S}$ be a smoothing operator in the power spectral domain, and let $\mathcal{M}$ denote an estimator for the background pseudo-stationary noise, derived using the MCRA approach [7]. The *transient beam-to-reference ratio* (TBRR) is defined by the ratio between the transient power of the beamformer output and the transient power of the strongest reference signal:

$$\Omega(k,\ell) = \frac{\mathcal{S}Y(k,\ell) - \mathcal{M}Y(k,\ell)}{\max\limits_{2\le i\le M}\left\{\mathcal{S}U_i(k,\ell) - \mathcal{M}U_i(k,\ell)\right\}}\,. \tag{4}$$

This ratio indicates whether a transient is likely derived from speech or ambient noise. Assuming that the steering error of the beamformer is relatively low, and that the interfering noise is uncorrelated with the desired speech, the TBRR is high only when speech components are present. Hence, by modifying the speech presence probability based on that ratio, we can generate a double mechanism for non-stationary noise reduction: First, through a fast update of the noise estimate (an increase in the noise estimate essentially results in lower spectral gain). Second, through the spectral gain computation (the spectral gain is exponentially modified by the speech presence probability [6]).

Let the *a priori* speech absence probability be estimated by

$$\hat{q}(k,\ell) = \begin{cases} 1, & \text{if } \Lambda(k,\ell) \le \Lambda_0 \text{ or } \Omega(k,\ell) < \Omega_{low} \\ \max\left\{\frac{\Omega_{high}-\Omega(k,\ell)}{\Omega_{high}-\Omega_{low}}, 0\right\}, & \text{otherwise,} \end{cases} \tag{5}$$

where $\Lambda(k,\ell) \triangleq \mathcal{S}Y(k,\ell)/\mathcal{M}Y(k,\ell)$ represents a local non-stationarity measure, Λ_0 denotes a threshold for detecting transients, and Ω_{low} and Ω_{high} are constants that represent the uncertainty in $\Omega(k,\ell)$ during weak speech activity. Based on a Gaussian statistical model [4], the speech presence probability is given by

$$p(k,\ell) = \left\{1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell))\exp(-\upsilon(k,\ell))\right\}^{-1} \tag{6}$$

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, and $\upsilon \triangleq \gamma\,\xi/(1+\xi)$. Computing a time-varying frequency-dependent smoothing parameter by $\tilde{\alpha}_d(k,\ell) = \alpha_d + (1-\alpha_d)\,p(k,\ell)$, we obtain the following estimate for the noise spectrum at the beamformer primary output:

$$\hat{\lambda}_d(k,\ell+1) = \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) + \beta[1-\tilde{\alpha}_d(k,\ell)]|Y(k,\ell)|^2 \tag{7}$$

where α_d ($0 < \alpha_d < 1$) represents the minimal value of the smoothing parameter, and β is a factor that compensates the bias when speech is absent. This estimate is fed into the OM-LSA estimator, for estimating the desired speech component of the beamformer output, $\hat{X}_1(k,\ell)$.

The proposed noise estimator takes into account transient, as well as stationary, noise components. When speech is absent, the TBRR is low. Consequently, $\hat{q}(k,\ell)$ increases to one, the speech presence probability decreases to zero, and $\tilde{\alpha}_d(k,\ell)$ reduces to its minimal value α_d. Hence, the fast update of the noise estimate and the low value of the speech presence probability facilitate the suppression of noise components, whether stationary or not. In case the noise abruptly changes

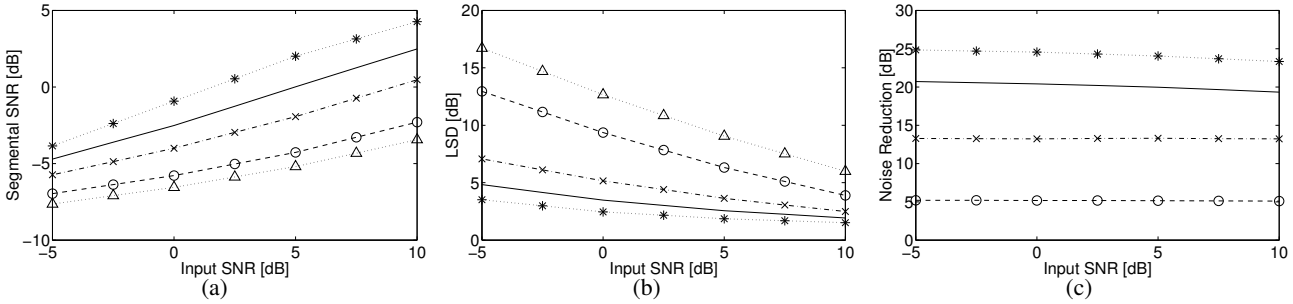

Fig. 2. (a) Segmental SNR, (b) log-spectral distance, and (c) noise reduction, at ($\triangle$) microphone #1, ($\circ$) TF-GSC output, ($\times$) single-channel postfiltering output, (solid line) multichannel postfiltering output, and ($*$) theoretical limit postfiltering output.

shortly before the onset of speech, single-channel noise estimation techniques either underestimate or overestimate the noise, depending whether it rises or decreases prior to the speech onset. The underestimation of the noise results in musical residual noise, while overestimation results in degradation of speech quality. On the other hand, the proposed method provides improved noise tracking capability allowing to better preserve weak speech components, while avoiding the musical residual noise phenomena.

4. Experimental Results

To validate the usefulness of the proposed multi-microphone LSA estimation approach under non-stationary noise conditions, we compare its performance to an alternative method in a noisy car environment. Specifically, multi-microphone speech signals are degraded by an interfering speaker and car noise signals. Then, TF-GSC beamforming is applied to the noisy signals, followed by either single-channel or multi-channel postfiltering. A linear array, consisting of four microphones with 5 cm spacing, is mounted in a car on the visor. Clean speech signals are recorded at a sampling rate of 8 kHz in the absence of background noise (standing car, silent environment). An interfering speaker and car noise signals are recorded while the car speed is about 60 km/h, and the window next to the driver is slightly open (about 5 cm; the other windows are closed). The input microphone signals are generated by mixing the speech and noise signals at various SNR levels in the range $[-5, 10]$ dB. Three objective quality measures are used: Segmental SNR (SegSNR), log-spectral distance (LSD), and noise reduction (NR) [9].

Fig. 2 shows experimental results obtained for various noise levels. The quality measures are evaluated at the first microphone, the adaptive beamformer output, and the postfiltering outputs. A theoretical limit postfiltering, achievable by calculating the noise spectrum from the noise itself, is also considered. Clearly, beamforming alone does not provide sufficient noise reduction in a car environment, owing to its limited ability to reduce diffuse noise [8]. Furthermore, multi-channel postfiltering is considerably better than single-channel postfiltering. The TF-GSC output is characterized by a high level of noise, which varies substantially due to the residual interfering components of speech, wind blows, and passing cars. Single-channel postfiltering suppresses the pseudo-stationary noise components, but is inefficient at attenuating the transient noise. By contrast, the proposed estimation approach achieves superior noise attenuation, while preserving the desired speech components. This is verified by a subjective study of speech spectrograms and informal listening tests.

5. 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. Speech and Audio Processing*, vol. 6, pp. 240–259, May 1998.

[3] K. U. Simmer, J. Bitzer, and C. Marro, *Post-Filtering Techniques*, pp. 39–60, In Brandstein and Ward [1], 2001.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-32, pp. 1109–1121, December 1984.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[6] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, October 2001.

[7] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, January 2002.

[8] 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.

[9] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.