TWO-CHANNEL SIGNAL DETECTION AND SPEECH ENHANCEMENT BASED ON THE TRANSIENT BEAM-TO-REFERENCE RATIO

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
http://www.AndreaElectronics.com

ABSTRACT

In reverberant and noisy environments, multichannel systems are designed for spatially filtering interfering signals coming from undesired directions. In case of incoherent or diffuse noise fields, beamforming alone does not provide sufficient noise reduction, and post-filtering is normally required. In this paper, we present a two-channel post-filtering approach for signal detection and speech enhancement. A mild assumption is made, that a desired signal component is stronger at the beamformer output than at the reference noise signal, and a noise component is stronger at the reference signal. The ratio between the transient power at the beamformer output and the transient power at the reference noise signal is used for indicating whether such a transient is desired or interfering. Experimental results demonstrate the usefulness of the proposed approach in a car environment.

1. INTRODUCTION

In reverberant and noisy environments, multichannel systems are designed for spatially filtering interfering signals coming from undesired directions [1]. In case of incoherent or diffuse noise fields, beamforming alone does not provide sufficient noise reduction, and post-filtering is normally required [2, 3, 4].

In this paper, we present a two-channel signal detection and speech enhancement approach based on the transient beam-to-reference ratio. A desired signal component is presumably stronger at the beamformer output than at the reference noise signal, and a noise component is stronger at the reference signal. Hence, the ratio between the transient power at beamformer output and the transient power at the reference signal indicates whether such a transient is desired or interfering. Based on a Gaussian statistical model [5], 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 for 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 mean-square error of the log-spectra. The performance of the proposed approach is evaluated in non-stationary car noise conditions. We demonstrate that single-channel post-filtering is inefficient at attenuating highly non-stationary noise components, since it lacks the ability to differentiate such components from the desired source components. By contrast, the proposed two-channel post-filtering approach achieves a significantly reduced level of background noise, whether stationary or not, without further distorting the signal components.

The paper is organized as follows. In Section 2, we review the two-channel generalized sidelobe canceller, and derive relations in the power-spectral domain between the beamformer output, the reference noise signal, the desired source signal, and the input transient interferences. In Section 3, the problem of signal detection in the time-frequency plane is addressed. In Section 4, we introduce an estimator for the time-varying spectrum of the beamformer output noise, and describe the two-channel speech enhancement approach.

2. TWO-CHANNEL GENERALIZED SIDELOBE CANCELLING

Let $x(t)$ denote a desired source signal, and let $d_{is}(t)$ and $d_{it}(t)$ denote uncorrelated interfering signals corresponding to the i-th sensor ($i = 1, 2$). The signal $d_{is}(t)$ represents the pseudo-stationary interferences, and $d_{it}(t)$ represents the undesired transient components. Assuming that the array is presteered to the direction of the source signal, the observed signals are given by

$$z_i(t) = x(t) + d_{is}(t) + d_{it}(t)\,, \quad i = 1, 2\,. \tag{1}$$

Using the short-time Fourier transform (STFT), we have in the time-frequency domain

$$Z_i(k,\ell) = X(k,\ell) + D_{is}(k,\ell) + D_{it}(k,\ell)\,, \tag{2}$$

where k represents the frequency bin index, and ℓ the frame index. Fig. 1 shows a two-channel generalized sidelobe canceller structure for a linearly constrained adaptive beamformer [8]. The beamformer comprises a fixed beamformer, a blocking channel which yields the reference noise signal $U(k,\ell)$, and an adaptive noise canceller $H(k,\ell)$ which eliminates 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 signal components may pass through the blocking channel due to steering error. The steering error is represented by

$$\Delta_k = \frac{\omega_k\, l}{c}\sin(\varphi) + \phi \tag{3}$$

Proc. ICASSP-2003, Hong Kong, 6–10 April 2003.

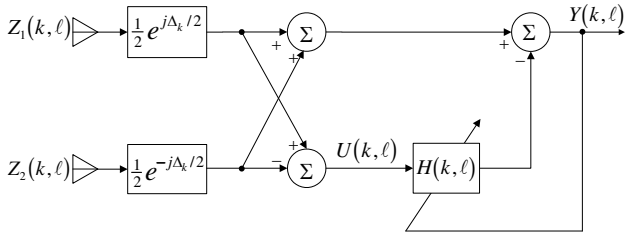

Fig. 1. Two-channel Generalized Sidelobe Canceller.

where ω_k is the center of the kth frequency bin, l is the distance between the sensors, c the speed of sound, φ the mismatch in the source direction, and ϕ the estimation error in the difference of phase.

Assuming homogeneous noise fields, the power-spectral density (PSD) matrices of the input noise signals are related to the corresponding spatial coherence functions, $\Gamma_s(k,\ell)$ and $\Gamma_t(k,\ell)$, by

$$\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell) = \lambda_s(k,\ell)\begin{bmatrix} 1 & \Gamma_s(k,\ell) \\ \Gamma_s^*(k,\ell) & 1 \end{bmatrix} \quad (4)$$

$$\mathbf{\Phi}_{\mathbf{D}_t\mathbf{D}_t}(k,\ell) = \lambda_t(k,\ell)\begin{bmatrix} 1 & \Gamma_t(k,\ell) \\ \Gamma_t^*(k,\ell) & 1 \end{bmatrix} \quad (5)$$

where $\lambda_s(k,\ell)$ and $\lambda_t(k,\ell)$ represent the input noise power at a single sensor. In this case, the optimal noise canceller, obtained by minimizing the output power of the stationary noise, is given by [9]

$$H(k,\ell) = \frac{j\Im\left\{e^{j\Delta_k}\Gamma_s(k,\ell)\right\}}{1-\Re\left\{e^{j\Delta_k}\Gamma_s(k,\ell)\right\}}\,. \quad (6)$$

Since the source signal, the stationary noise and transient noise are uncorrelated, the input PSD-matrix is given by

$$\mathbf{\Phi}_{\mathbf{ZZ}}(k,\ell) = \lambda_x(k,\ell)\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell) + \mathbf{\Phi}_{\mathbf{D}_t\mathbf{D}_t}(k,\ell) \quad (7)$$

where $\lambda_x(k,\ell) \triangleq E\left\{|X(k,\ell)|^2\right\}$ is the PSD of the desired source signal. Using (4)–(7) and Fig. 1, we obtain the following linear relations between the PSD's of the beamformer output, the reference signal, the desired source signal, and the input interferences:

$$\begin{aligned}\phi_{YY}(k,\ell) &= C_{11}(k,\ell)\lambda_x(k,\ell) + C_{12}(k,\ell)\lambda_s(k,\ell) \\ &\quad + C_{13}(k,\ell)\lambda_t(k,\ell) \quad (8)\\ \phi_{UU}(k,\ell) &= C_{21}(k)\lambda_x(k,\ell) + C_{22}(k,\ell)\lambda_s(k,\ell) \\ &\quad + C_{23}(k,\ell)\lambda_t(k,\ell) \quad (9)\end{aligned}$$

where

$$C_{11}(k,\ell) = \left[\cos\left(\tfrac{\Delta_k}{2}\right) - \frac{\Im\left\{e^{j\Delta_k}\Gamma_s(k,\ell)\right\}\sin\left(\frac{\Delta_k}{2}\right)}{1-\Re\left\{e^{j\Delta_k}\Gamma_s(k,\ell)\right\}}\right]^2 \quad (10)$$

$$C_{12}(k,\ell) = \frac{1-|\Gamma_s(k,\ell)|^2}{1-\Re\left\{e^{j\Delta_k}\Gamma_s(k,\ell)\right\}} \quad (11)$$

$$\begin{aligned}C_{13}(k,\ell) &= \frac{1}{2}\left|1+H(k,\ell)\right|^2 \\ &\quad + \frac{1}{2}\Re\left\{e^{j\Delta_k}\Gamma_t(k,\ell)\left[1+H(k,\ell)\right]^2\right\} \quad (12)\end{aligned}$$

$$C_{21}(k) = \sin^2\left(\tfrac{\Delta_k}{2}\right) \quad (13)$$

$$C_{22}(k,\ell) = \frac{1}{2}\left[1-\Re\left\{e^{j\Delta_k}\Gamma_s(k,\ell)\right\}\right] \quad (14)$$

$$C_{23}(k,\ell) = \frac{1}{2}\left[1-\Re\left\{e^{j\Delta_k}\Gamma_t(k,\ell)\right\}\right]\,. \quad (15)$$

3. SIGNAL DETECTION

Transient *signal* components are relatively strong at the beamformer output, whereas transient *noise* components are relatively strong at the reference signal. Hence, the transient power ratio between the beamformer output and the reference signal is expected to be large for desired transients, and small for noise components. Let $\mathcal{S}$ be a smoothing operator in the power spectral domain,

$$\mathcal{S}Y(k,\ell) = \alpha_s\cdot\mathcal{S}Y(k,\ell-1) + (1-\alpha_s)\sum_{i=-w}^{w} b_i|Y(k-i,\ell)|^2 \quad (16)$$

where α_s ($0\le\alpha_s\le 1$) is a parameter for the smoothing in time, and b is a normalized window function ($\sum_{i=-w}^{w} b_i = 1$) that determines the smoothing in frequency. Let $\mathcal{M}$ denote an estimator for the PSD of the background pseudo-stationary noise, derived using the MCRA approach [7]. We define the *transient beam-to-reference ratio* (TBRR) by the ratio between the transient power of the beamformer output and the transient power of the reference signal:

$$\Omega(k,\ell) = \frac{\mathcal{S}Y(k,\ell)-\mathcal{M}Y(k,\ell)}{\mathcal{S}U(k,\ell)-\mathcal{M}U(k,\ell)}\,. \quad (17)$$

Let three hypotheses H_{0s}, H_{0t}, and H_1 indicate respectively absence of transients, presence of an interfering transient, and presence of a desired transient at the beamformer output. Then, given that H_1 or H_{0t} is true, we have

$$\begin{aligned}\Omega(k,\ell)|_{H_1\cup H_{0t}} &\approx \frac{\phi_{YY}(k,\ell)-C_{12}(k,\ell)\lambda_s(k,\ell)}{\phi_{UU}(k,\ell)-C_{22}(k,\ell)\lambda_s(k,\ell)} \\ &= \frac{C_{11}(k,\ell)\lambda_x(k,\ell)+C_{13}(k,\ell)\lambda_t(k,\ell)}{C_{21}(k)\lambda_x(k,\ell)+C_{23}(k,\ell)\lambda_t(k,\ell)}\,. \quad (18)\end{aligned}$$

Assuming there exist thresholds $\Omega_{\mathrm{high}}(k)$ and $\Omega_{\mathrm{low}}(k)$ such that

$$\begin{aligned}\Omega(k,\ell)|_{H_{0t}} &\approx \frac{C_{13}(k,\ell)}{C_{23}(k,\ell)} \le \Omega_{\mathrm{low}}(k) \le \Omega_{\mathrm{high}}(k) \\ &\le \frac{C_{11}(k,\ell)}{C_{21}(k)} \approx \Omega(k,\ell)|_{H_1} \quad (19)\end{aligned}$$

we determine the likelihood of signal presence proportionally to $\Omega(k,\ell)$ by

$$\psi(k,\ell) = \begin{cases} 0\,, & \text{if } \Omega(k,\ell)\le\Omega_{\mathrm{low}}(k) \\ 1\,, & \text{if } \Omega(k,\ell)>\Omega_{\mathrm{high}}(k) \\ \frac{\Omega(k,\ell)-\Omega_{\mathrm{low}}(k)}{\Omega_{\mathrm{high}}(k)-\Omega_{\mathrm{low}}(k)}\,, & \text{otherwise.}\end{cases} \quad (20)$$

The decision rules for detecting transients at the beamformer output and reference signal are

$$\Lambda_Y(k,\ell) \triangleq \mathcal{S}Y(k,\ell)/\,\mathcal{M}Y(k,\ell) > \Lambda_0 \quad (21)$$

$$\Lambda_U(k,\ell) \triangleq \mathcal{S}U(k,\ell)/\,\mathcal{M}U(k,\ell) > \Lambda_0\,, \quad (22)$$

respectively, where Λ_Y and Λ_U denote measures of the local non-stationarities (LNS) [9]. For a given signal, the LNS fluctuates

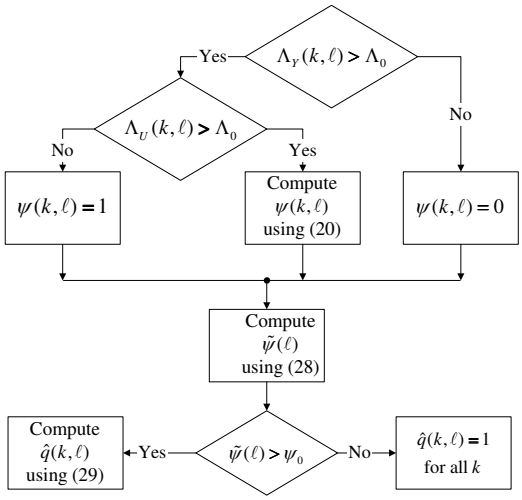

Fig. 2. Block diagram for the *a priori* signal absence probability estimation.

about one in the absence of transients, and increases well above one in the neighborhood of time-frequency bins that contain transients. The false alarm and detection probabilities are defined by

$$P_{f,Y}(k,\ell) = \mathcal{P}\left(\Lambda_Y(k,\ell) > \Lambda_0 \mid H_{0s}\right) \quad (23)$$

$$P_{d,Y}(k,\ell) = \mathcal{P}\left(\Lambda_Y(k,\ell) > \Lambda_0 \mid H_1 \cup H_{0t}\right). \quad (24)$$

Then for a specified $P_{f,Y}$, the required threshold value and the detection probability are given by [9]

$$\Lambda_0 = \frac{1}{\mu} F^{-1}_{\chi^2;\mu}\left(1 - P_{f,Y}\right) \quad (25)$$

$$P_{d,Y}(k,\ell) = 1 - F_{\chi^2;\mu}\left[\frac{1}{1+\xi_Y(k,\ell)} F^{-1}_{\chi^2;\mu}\left(1 - P_{f,Y}\right)\right] \quad (26)$$

where

$$\xi_Y(k,\ell) \triangleq \frac{C_{11}(k,\ell)\lambda_x(k,\ell) + C_{13}(k,\ell)\lambda_t(k,\ell)}{C_{12}(k,\ell)\lambda_s(k,\ell)} \quad (27)$$

represents the ratio between the transient and pseudo-stationary power at the beamformer output, and $F_{\chi^2;\mu}(x)$ denotes the standard chi-square distribution function with μ degrees of freedom.

To improve the discrimination between *wideband* source and interfering transients, we also compute for each frame a global likelihood of signal presence. The global likelihood is related to the number of frequency bins that likely contain desired components within a certain frequency range:

$$\tilde{\psi}(\ell) = \frac{1}{k_1 - k_0 + 1} \sum_{k=k_0}^{k_1} \psi(k,\ell) \quad (28)$$

where k_0 and k_1 are the lower and upper frequency bin indices representing the frequency range. The global likelihood is compared to a certain threshold ψ_0. In case the global likelihood is too low, we conclude that signal is absent from that frame and set the *a priori* signal absence probability to one for all frequency bins. This prevents from narrow-band interfering transients, particularly those arriving from the look direction, to be confused with desired components. This also helps to reduce musical noise phenomena.

$\Lambda_0 = 1.54$	$\Omega_{\text{low}} = 1$	$\Omega_{\text{high}} = 3$	$\gamma_0 = 4.6$
$\alpha = 0.92$	$\alpha_s = 0.8$	$\alpha_d = 0.85$	$\beta = 1.98$
$k_0 = 9$	$k_1 = 113$	$\psi_0 = 0.25$	$\mu = 22.1$
$b = \frac{1}{12}[1 \quad 3 \quad 4 \quad 3 \quad 1]$	$N = 256$	$G_{min} = -20$ dB	

Table 1. Values of parameters used in the implementation of the two-channel post-filtering, for a sampling rate of 8 kHz

Fig. 2 summarizes a block diagram for the estimation of the *a priori* signal absence probability. First we detect transients at the beamformer output and reference signal. The likelihood of signal presence $\psi(k,\ell)$ is set to zero if no transients are detected at the beamformer output, and set to one if a transient is detected at the beamformer output but not at the reference signal. In case a transient is detected simultaneously at the beamformer output and at the reference signal, $\psi(k,\ell)$ is computed via (20). Subsequently, a global likelihood $\tilde{\psi}(\ell)$ is generated, and compared to the threshold ψ_0. In case the global likelihood is above the threshold, the *a priori* signal absence probability is determined by the *a posteriori* SNR at the beamformer output with respect to the pseudo-stationary noise $\gamma_s(k,\ell) \triangleq |Y(k,\ell)|^2 / \mathcal{M}Y(k,\ell)$:

$$\hat{q}(k,\ell) = \begin{cases} 1, & \text{if } \gamma_s(k,\ell) \le 1 \text{ or } \tilde{\psi}(\ell) \le \psi_0 \\ \max\left\{\frac{\gamma_0 - \gamma_s(k,\ell)}{\gamma_0 - 1}, 1 - \psi(k,\ell)\right\}, & \text{otherwise,} \end{cases} \quad (29)$$

where γ_0 is a constant satisfying $\mathcal{P}\left(\gamma_s(k,\ell) \ge \gamma_0 \mid H_{0s}\right) < \epsilon$ for a certain significance level ϵ (typically $\epsilon = 0.01$ and $\gamma_0 = -\log(\epsilon) = 4.6$).

4. SPEECH ENHANCEMENT

In this section, we estimate the spectrum of the beamformer output noise and the clean signal based on the MCRA and the *optimally-modified log-spectral amplitude* (OM-LSA) gain function.

The MCRA approach for noise spectrum estimation 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

$$\hat{\lambda}_d(k,\ell+1) = \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) + \beta \cdot [1 - \tilde{\alpha}_d(k,\ell)]|Y(k,\ell)|^2 \quad (30)$$

where $\lambda_d(k,\ell)$ is the total noise PSD at the beamformer output, $\tilde{\alpha}_d(k,\ell)$ is a time-varying frequency-dependent smoothing parameter, and β is a factor that compensates the bias when signal is absent. The smoothing parameter is determined by the signal presence probability, $p(k,\ell)$, and a constant α_d ($0 < \alpha_d < 1$) that represents its minimal value:

$$\tilde{\alpha}_d(k,\ell) \triangleq \alpha_d + (1 - \alpha_d)\, p(k,\ell)\,. \quad (31)$$

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

Based on a Gaussian statistical model [5], the signal presence

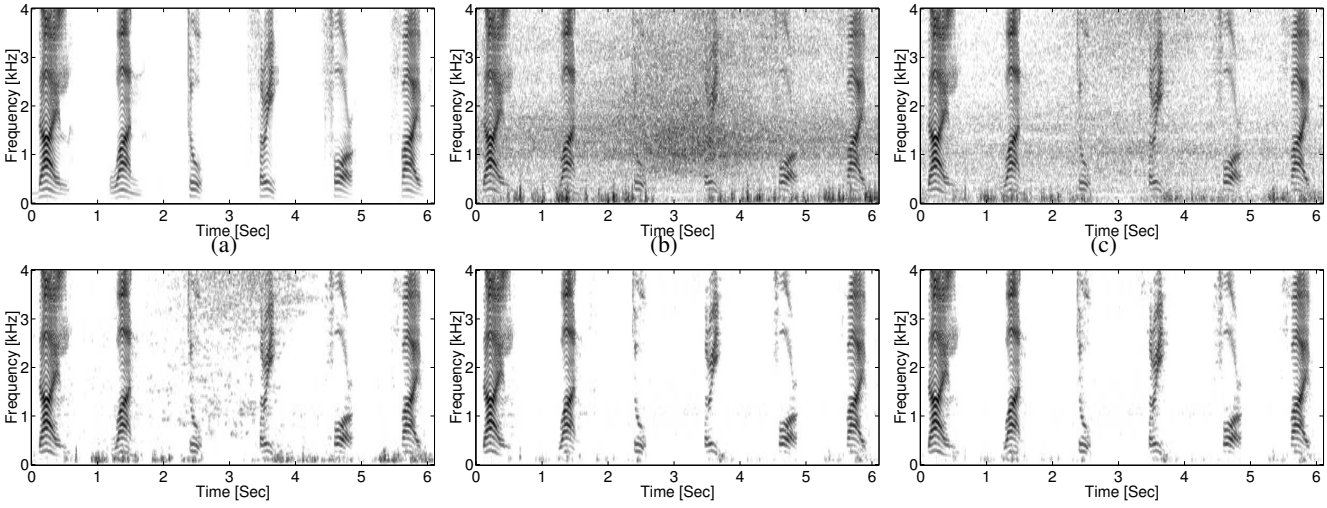

Fig. 3. Speech spectrograms. (a) Original clean speech signal at microphone #1: "Dial one two three four five."; (b) Noisy signal at microphone #1 (SegSNR $= -6.5$ dB); (c) Beamformer output (SegSNR $= -5.0$ dB); (d) Single-channel post-filtering output (SegSNR $= -3.0$ dB); (e) Two-channel post-filtering output (SegSNR $= -0.9$ dB); (f) Theoretical limit (SegSNR $= -0.5$ dB).

probability is given by

$$p(k,\ell) = \left\{1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell))\exp(-\upsilon(k,\ell))\right\}^{-1} \tag{32}$$

where $\xi(k,\ell) \triangleq \lambda_x(k,\ell)/\lambda_d(k,\ell)$ is the *a priori* SNR, $\upsilon(k,\ell) \triangleq \gamma(k,\ell)\cdot\xi(k,\ell)/(1+\xi(k,\ell))$, and $\gamma(k,\ell) \triangleq |Y(k,\ell)|^2/\lambda_d(k,\ell)$ is the *a posteriori* SNR. The *a priori* SNR is estimated by [6]

$$\hat{\xi}(k,\ell) = \alpha G_{H_1}^2(k,\ell-1)\gamma(k,\ell-1) + (1-\alpha)\max\{\gamma(k,\ell)-1, 0\} \tag{33}$$

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

$$G_{H_1}(k,\ell) \triangleq \frac{\xi(k,\ell)}{1+\xi(k,\ell)}\exp\left(\frac{1}{2}\int_{\upsilon(k,\ell)}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{34}$$

is the spectral gain function of the *Log-Spectral Amplitude* (LSA) estimator when signal is surely present. The estimate for the clean signal STFT is given by

$$\hat{X}(k,\ell) = G(k,\ell)Y(k,\ell)\,, \tag{35}$$

where

$$G(k,\ell) = \{G_{H_1}(k,\ell)\}^{p(k,\ell)}\cdot G_{\min}^{1-p(k,\ell)} \tag{36}$$

is the OM-LSA gain function and $G_{\min}$ denotes a lower bound constraint for the gain when signal is absent. Typical parameter values for a sampling rate of 8 kHz, are given in Table 1.

To validate the usefulness of the proposed approach under non-stationary noise conditions, we compare its performance to a single-channel post-filtering in various car environments. Specifically, two-channel speech signals are degraded by interfering speakers and various car noise types. Then, beamforming is applied to the noisy signals, followed by either single-channel or two-channel post-filtering. A theoretical limit post-filtering, achievable by calculating the noise spectrum from the noise itself, is also considered. Typical examples of speech spectrograms are presented in Fig. 3. The window next to the driver is slightly open, inducing transient low-frequency noise due to wind blows, and wideband transient noise due to passing cars. The beamformer output is characterized by a high level of noise, owing to its limited ability to reduce diffuse noise. Its enhancement using single-channel post-filtering well suppresses the pseudo-stationary noise, but adversely retains the transient noise components. By contrast, the enhancement using two-channel post-filtering results in superior noise attenuation, while preserving the desired source components.

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. SAP*, vol. 6, no. 1, pp. 240–259, May 1998.

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

[4] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. ICASSP-2002*, pp. 901–904.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[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] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. AP-30, no. 1, pp. 27–34, January 1982.

[9] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering," Tech. Rep., EE Pub. 1332, Technion - IIT, Haifa, Israel, July 2002.