MULTI-CHANNEL PSYCHOACOUSTICALLY MOTIVATED SPEECH ENHANCEMENT

Justinian Rosca, Radu Balan

Siemens Corporate Research
Multimedia and Video Technology
755 College Road East
Princeton, NJ 08540
Justinian.Rosca,Radu.Balan@scr.siemens.com

Christophe Beaugeant

Siemens AG
Information and Communication Mobile
Grillparzerstr. 10-18
81737 Munich, Germany
Christophe.Beaugeant@mch.siemens.de

ABSTRACT

Multichannel techniques offer advantages in noise reduction and overall output signal quality when compared to the well studied mono approaches. In this paper we present an original multichannel psychoacoustically motivated noise reduction algorithm that naturally extends the single channel psychoacoustic masking filter previously studied in the literature [1]. The optimality criterion is designed to simultaneously satisfy the psychoacoustic masking principle and minimize the signal total distortion. In experiments on real data recorded in a noisy car environment, we show the enhanced performance of the two-channel solution in terms of artifacts and overall tradeoff between artefacts and amount of noise removed as given by word recognition rates.

1. INTRODUCTION

Recent signal processing literature describes a variety of approaches to the ubiquitous noise reduction / speech enhancement problem. Many approaches still use a single microphone solution ([2] and references therein). Research in microphone-array systems indicate the promise of such techniques for speech enhancement and hands-free communication applications in noisy environments [3, 4]. Theoretically, multi-channel techniques offer more information about the acoustic environment, therefore should indeed offer possibility for improvement especially in the case of reverberant environments due to multi-path effects and severe noise conditions known to affect the performance of state-of-the-art single channel techniques. The effectiveness of multiple channel techniques for just a few microphones is yet to be proven.

Beamforming techniques and in general approaches grounded in the array processing literature [5, 6] indicate tiny SNR improvements in the case of a small number of microphones. Rather than directly considering such approaches, we explore extensions of successful mono noise reduction technique to multiple channels.

Outstanding among the mono approaches is the psychoacoustically motivated method proposed in [7]. This method uses an observation from human hearing studies known as tonal masking. Tonal masking means that a given tone becomes inaudible by a listener if another tone (the masking tone) with a similar or slightly different frequency is simultaneously presented to the listener [8]. This means that for a given speech signal (or more specific, for a given spectral power density), there is a psychoacoustic spectral threshold so that any interferer of spectral power below this threshold becomes unnoticed.

Most denoising schemes trade off speech intelligibility (for instance as measured by the articulation index [9]) for the amount of noise removed as measured by signal-to-noise-ratio (SNR) [7]. Moreover, it is sometimes desirable to preserve the background noise characteristics. Therefore the entire removal of the noise is neither desirable nor possible. More feasible is to reduce noise level down to the psychoacoustic threshold level but not below it. In this framework, our approach generalizes the psychoacoustic noise reduction approach to multiple channels. Our approach exploits the microphone array signals to further enhance the useful speech signal at reduced level of artifacts.

The layout of the paper is as follows. Next section introduces the model assumptions used in the present approach. Section 3 derives the psychoacoustic masking based signal estimator. Then section 4 states the identification algorithms for the components of our model. Section 5 discusses three evaluation criteria used (objective, subjective and speech recognition rates) and presents experimental results on noisy car speech data. We compare the two-channel speech enhancement system with the single channel psychoacustic method of [1]. Section 6 summarizes this work.

2. MULTI-CHANNEL MODEL AND ASSUMPTIONS

The mixing model we consider is as follows. We assume D microphone signals $x_1(t), \ldots, x_D(t)$, record a source $s(t)$ and noise $n_1(t), \ldots, n_D(t)$ signals,

$$x_l(t) = \sum_{k=0}^{L_l} a_k^l s(t - \tau_k^l) + n_l(t) \tag{1}$$

where (a_k^l, τ_k^l) are the attenuation and delay on the k^{th} path to microphone l. In frequency domain, the convolutions become multiplications. Furthermore, since we are not interested in balancing the channels, we redefine the source so that the first channel becomes unity:

$$\begin{array}{rcl} X_1(k,\omega) & = & S(k,\omega) + N_1(k,\omega) \\ X_2(k,\omega) & = & K_2(\omega)S(k,\omega) + N_2(k,\omega) \\ & \cdots & \\ X_D(k,\omega) & = & K_D(\omega)S(k,\omega) + N_D(k,\omega) \end{array} \tag{2}$$

where k is the frame index, and ω the frequency index. More compactly, this model can be rewritten as

$$X = KS + N \tag{3}$$

where X, K, S, N are D-complex vectors. Regarding this model, we make the following assumptions:

1. $S(\omega)$ are zero-mean stochastic processes with spectral power $\rho_s(\omega) = \mathbf{E}[|S|^2]$;

2. $(N_1, N_2, \ldots, N_D)$ is a zero-mean stochastic signal with spectral covariance matrix

$$R_n(\omega) = \begin{bmatrix} \mathbf{E}[|N_1|^2] & \mathbf{E}[N_1\overline{N_2}] & \cdots & \mathbf{E}[N_1\overline{N_D}] \\ \mathbf{E}[N_2\overline{N_1}] & \mathbf{E}[|N_2|^2] & \cdots & \mathbf{E}[N_2\overline{N_D}] \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{E}[N_D\overline{N_1}] & \mathbf{E}[N_D\overline{N_2}] & \cdots & \mathbf{E}[|N_D|^2] \end{bmatrix}; \tag{4}$$

3. s is independent of n.

Section 4 describes how we estimate K, ρ_s and R_n.

3. ALGORITHM DESIGN

Consider a linear filter

$$A = \begin{bmatrix} A_1 & A_2 & \cdots & A_D \end{bmatrix} \tag{5}$$

applied to the measured signals $X_1, \ldots, X_D$. The output becomes:

$$Y = \sum_{l=1}^{D} A_l X_l = AKS + AN \tag{6}$$

Suppose we would like to obtain an estimate of S that contains a small amount of noise. Let $0 \leq \zeta_1, \ldots, \zeta_D \leq 1$ be given constants such that the desired signal is:

$$W = S + \zeta_1 N_1 + \zeta_2 N_2 + \cdots + \zeta_D N_D \tag{7}$$

Since source signal s and noise signal n are independent, the estimation error $E = Y - W$ has the variance:

$$R_e = |AK - 1|^2 \rho_s + (A - \zeta) R_n (A^* - \zeta^T) \tag{8}$$

where $\zeta = [\zeta_1, \cdots, \zeta_D]$ is the $1 \times D$ vector of desired levels of noise. In [1], the idea was to design the filter(s) so that the distortion term due to noise achieves a preset value R_T, called the *threshold masking*, depending solely on the signal spectral power ρ_s. The idea is that any noise whose spectral power is below this threshold is unnoticed, therefore why cancel the noise completely when we don't notice it below R_T; furthermore, by doing less noise removal, the artifacts would be smaller as well (see detailed descrption of this in [1]). Following this idea, we would like our filter to achieve a noise distortion level of R_T. Yet, we have D unknowns and one constraint so far. This leaves us with $D-1$ degrees of freedom. We can use these degrees of freedom to choose A that minimizes the total distortion. Thus the optimization problem becomes:

$$argmin_A R_e \quad , \quad \text{subject to } (A - \zeta) R_n (A^* - \zeta^T) = R_T \tag{9}$$

Suppose A_o is the optimal solution. Then we validate it by checking whether $|A_o K| \leq 1$. When it is not true, we choose not to do any processing (perhaps the noise level is already lower than the threshold, so why amplify it). Hence:

$$A = \begin{cases} A_o & \text{if} \quad |A_o K| \leq 1 \\ (1, 0, \cdots, 0) & \text{otherwise} \end{cases} \tag{10}$$

Set $B = A - \zeta$ and construct the Lagrangian:

$$L(B, \lambda) = |BK + \zeta K - 1|^2 \rho_s + B R_n B^* + \lambda (B R_n B^* - R_T)$$

we obtain the system:

$$\begin{aligned} K^*(BK + \zeta K - 1)\rho_s + BR_n + \lambda B R_n &= 0 \\ K(K^* B^* + B^* \zeta^T - 1)\rho_s + R_n B^* + \lambda R_n B^* &= 0 \\ B R_n B^* - R_T &= 0 \end{aligned}$$

Solving for B in the first equation and inserting the expression into the second equation, we obtain with $\mu = (1 + \lambda)/\rho_s$:

$$R_T = |1 - \zeta K|^2 K^* (\mu R_n + KK^*)^{-1} R_n (\mu R_n + KK^*)^{-1} K$$

Using the Inversion Lemma (see [2]) the equation in μ becomes:

$$\mu = -K^* R_n^{-1} K \pm |1 - \zeta K| \sqrt{\frac{K^* R_n^{-1} K}{R_T}} \tag{11}$$

Replacing in R_e, we obtain:

$$R_e = R_T + \rho_s | \pm \sqrt{R_T (K^* R_n^{-1} K)} - |1 - \zeta K||^2$$

Hence the optimal solution is the one with '$+$' in (11). Consequently, the optimizer becomes:

$$A_o = \zeta + \frac{1 - \zeta K}{|1 - \zeta K|} \sqrt{\frac{R_T}{K^* R_n^{-1} K}} K^* R_n^{-1} \tag{12}$$

In practical application we have chosen $\zeta_1 = \zeta$ and $\zeta_k = 0, k > 1$, which simplifies the final expression of our estimator. Then:

$$A_o = (\zeta, 0, \cdots, 0) + \sqrt{\frac{R_T}{K^* R_n^{-1} K}} K^* R_n^{-1} \tag{13}$$

and

$$|A_o K| = \zeta + \sqrt{R_T (K^* R_n^{-1} K)}$$

Note that we have obtained a closed form solution for the filter A in equation 6, similar to the one obtained in [1]. On the other hand, the connexion between single-channel and multi-channel estimators is similar to connexion established in [10]. Indeed, the multidimensional signal X is first filtered by $\frac{1}{K^* R_n^{-1} K} K^* R_n^{-1}$ to a scalar signal

$$Z = \frac{1}{K^* R_n^{-1} K} K^* R_n^{-1} X = S + \frac{1}{K^* R_n^{-1} K} K^* R_n^{-1} N \tag{14}$$

whose noise effective spectral power given by $R_n^{eff} = \frac{1}{K^* R_n^{-1} K}$. Then, the Gustafsson filter $H_G = \zeta + \sqrt{\frac{R_T}{R_N^{eff}}}$ is applied to Z in order to obtain the estimation Y. Among many possible Multi-Input-Single-Output filter, the choice (14) is optimal with respect to criterion (9).

4. MODEL IDENTIFICATION

We turn now our attention to the estimation and identification of the model stated in the previous section.

4.1. An Adaptive Model-Based Estimator of K

In this work we consider what we call an adaptive *model-based* estimator of K, which makes use of the simple but effective direct-path mixing model. Accordingly, the transfer function ratios are parameterized by only delay and attenuation parameters:

$$K_l(\omega) = a_l e^{i\omega\delta_l} \ , \ l \geq 2 \tag{15}$$

The idea is to update the direct-path model while observing the mixing equations in order to track K adaptively.

Considering the statistical independence signal-noise, the short-time spectral power of the measured signal $R_x(k, \omega)$ is:

$$R_x(k, \omega) = \rho_s(k, \omega) KK^* + R_n(k, \omega) \tag{16}$$

Delay and attenuation parameters are determined so that we best fit (16), for every ω, in the Frobenius norm ($\|A\|_F^2 = trace\{AA^*\}$). Thus the criterion to be minimized is:

$$J(a_2,\dots,a_D,\delta_2,\dots,\delta_D) = \sum_\omega trace\{(R_x - R_n - \rho_s KK^*)^2\} \tag{17}$$

The summation across the frequencies shows that the same parameters $(a_l,\delta_l)_{2\le l\le D}$ have to explain all the frequencies. The gradient of J evaluated on the current estimate $(a_l,\delta_l)_{2\le l\le D}$ is

$$\frac{\partial J}{\partial a_l} = -4\sum_\omega \rho_s \cdot \mathrm{real}(K^* E v_l) \tag{18}$$

$$\frac{\partial J}{\partial \delta_l} = -2a_l\sum_\omega \omega\rho_s \cdot \mathrm{imag}(K^* E v_l) \tag{19}$$

where $E = R_x - R_n - \rho_s KK^*$ and $v_l = [0 \cdots 0 \; e^{i\omega\delta_l} \; 0 \cdots 0]^T$. In words, v_l is a D-vector of zeros everywhere except on the l^{th} entry where it is $e^{i\omega\delta_l}$.

Then the parameter update rules are:

$$a_l^{'} = a_l - \alpha\frac{\partial J}{\partial a_l} \tag{20}$$

$$\delta_l^{'} = \delta_l - \alpha\frac{\partial J}{\partial \delta_l} \tag{21}$$

where $0 \le \alpha \le 1$ is the learning rate.

4.2. Estimation of Spectral Power Densities

The estimation of the noise covariance R_n is based on a *voice activity detector* (VAD) signal:

$$R_n = \begin{cases} (1-\beta)R_n^{old} + \beta XX^* & \text{if voice present} \\ R_n^{old} & \text{otherwise} \end{cases} \tag{22}$$

An approximate ρ_s is satisfactory. [7] outlined that even a rough estimate by spectral subtraction is good enough in psychoacoustic filtering. ρ_s would not be used directly in the signal estimation Y (equation 6), but only in the evaluation of the masking threshold R_T and the update rules for delay and attenuation parameters (and therefore K). Therefore the signal spectral power ρ_s is estimated by spectral subtraction:

$$\rho_s = \begin{cases} R_{x;11} - R_{n;11} & \text{if} \quad R_{x;11} > \beta_{SS}R_{n;11} \\ (\beta_{SS}-1)R_{n;11} & \text{if} \quad \text{otherwise} \end{cases} \tag{23}$$

where $\beta_{SS} > 1$ is a noise floor constant.

4.3. Estimation of Voice Presence: VAD

For voice estimation we rely on the multi-channel approach described in [11] and exploit the spatial location of the voice source even in the presence of diffuse noise. The main idea is that a multi-channel filter H to maximize filter output SNR is extremely good at highlighting voice presence in adverse environments, although it may be poor for speech enhancement by itself. The maximum SNR objective criterion is:

$$J(H) = \frac{\mathbf{E}[|AKS|^2]}{\mathbf{E}[|AN|^2]} = \frac{\rho_s AKK^*A}{AR_nA^*} \tag{24}$$

[11] shows that the closed form solution for H can be obtained by solving a generalized eigenvalue problem:

$$H = \rho_s K^* R_n^{-1} \tag{25}$$

H above has the property of maximizing output energy when the voice signal is present. The resulting VAD is:

$$VAD(k) = \begin{cases} 1 & if \quad |HX|^2 \ge \gamma|X|^2 \\ 0 & otherwise \end{cases} \tag{26}$$

where $\gamma > 0$ is a constant boosting factor for the signal energy. Note that the VAD decision on the present data frame will only be used in the next frame estimation of noise parameters, and subsequently speech spectral power, transfer function ratios K and finally the adaptive filter A.

With the design of the VAD, now we have a complete scheme (see Figure 1) for multi-channel psycho-acoustically motivated speech enhancement, whose implementation we test next.

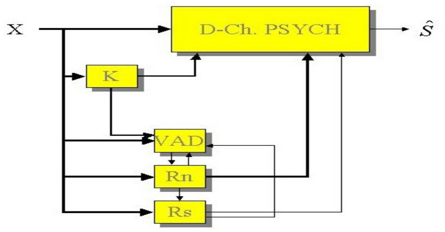

Fig. 1. Block diagram of the multi-channel psychoacoustic speech enhancement system.

5. EXPERIMENTAL RESULTS

For a practical implementation, we considered the case of two channels, $D = 2$. We used $8KHz$ stereo recordings from a noisy car environment. Input data had on average -6.5dB overall SNR. Example waveforms are plotted in Figure 2.

The time-frequency analysis was done with a Hamming window of size 512 samples with 50% overlap. R_x was estimated by a first-order filter with learning rate of 0.9. Other parameters were $\beta_{SS} = 1.1$ (see equation 23), $\beta = 0.2$ (22), $\zeta = 0.001$ (12), learning rate $\alpha = 0.01$ (21), VAD boosting factor $\gamma = 100$.

We applied the two-channel psychoacoustic noise reduction algorithm on a set of voices (male and female) superimposed with noise segments from two noise files (four combinations overall). For comparison, we also implemented the single channel psychoacoustic masking based filter as proposed in [7].

The evaluation of the speech enhancement algorithm consisted of (1) qualitative subjective measure; (2) objective measures; and (3) automatic speech recognition (word error rate measure). The qualitative measure is based on mean opinion score rankings and qualitative listening. The objective measures reported here are the average instantaneous SNR gain and distortion, each defined as follows:

$$aGain = \frac{1}{M}\sum_k 10\log_{10}\frac{\|AKS(k)\|^2}{\|S(k)\|^2}\frac{\|N_1(k)\|^2}{\|S_1(k)\|^2} \tag{27}$$

$$Dist = \frac{1}{M}\sum_k 10\log_{10}\frac{\|y_s - s_1\|}{\|s_1\|} \tag{28}$$

where M is the number of blocks when voice is present and y_s is the time domain estimate of the signal component in Y (that is AKS). Ideally, the distortion measure should be a large negative number while the average SNR gain should be a large positive number. The detailed description of the automatic speech recognition implementation, experiments and evaluation is provided in [12].

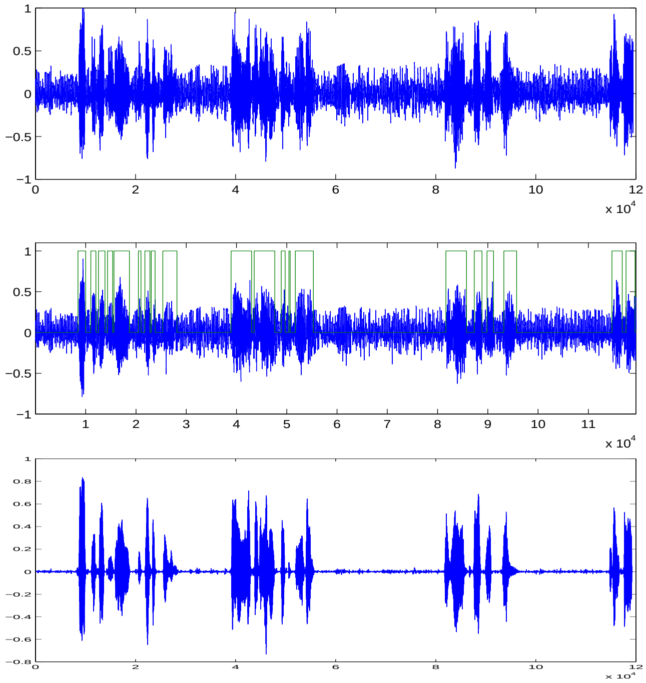

Fig. 2. Input 2-channel waveforms (top two plots), and the filter output (bottom); the VAD decision is superimposed on the second channel waveform.

Two Channel		Single Channel	
aGain	Dist	aGain	Dist
3.83	-15.35	6.58	-8.94
2.72	-21.56	6.14	-10.68
5.35	-20.38	7.81	-18.19
4.57	-22.93	7.41	-20.75
4.12	-20.06	6.99	-14.64

Table 1. Comparison of objective measures: average instantaneous SNR gain (aGain) and distortion (Dist), for two channel and one channel psychoacoustic algorithms, four data sets.

Tables 1 and 2 present the results obtained. In terms of average instantaneous gain, Table 1 shows better SNR numbers for the single channel solution, but considerably lower distortion for the two channel algorithm versus the single channel algorithm. Listening tests strengthened this conclusion. Furthermore, speech recognition tests showed improved word recognition rates particularly under high noise conditions. The mono speech enhancement alternative in the automatic speech recognition tests was recursive Spectral Subtraction (see [12]), which is similar to our speech spectral power estimation.

The two channel algorithm output had little speech distortion and noise artifacts compared to the mono solution, being clearly the preferred choice.

6. CONCLUSION

Scheme	Clean speech SNR $\geq$ 20 dB	Low car noise $7 \leq$ SNR ≤ 14 dB	High car noise SNR ≤ 7 dB
Ch.1 (in)	80.9%	55.0%	30.15%
Ch.2 (in)	86.93%	68.5%	35.67%
Ch.1 (SS)	88.44%	73.50%	63.81%
Ch.2 (SS)	90.95%	75.00%	51.76%
Two-ch.	92.69%	94.5%	77.38%

Table 2. Word recognition rates for data at three noise levels: (1) unprocessed; (2) processed by a mono spectral subtraction (3) enhanced by the two-channel psychoacoustic algorithm.

Taking into account psychoacoustic masking principles for speech enhancement has proven useful on sigle channel data. In this paper we extend the approach to the multi-channel case. The solution is original in its extension of the single channel psychoacoustically motivated constraint. The optimality criterion not only satisfies this principle but also minimizes the total signal distortion. The role of the latter constraint is critical: significantly reduced distortions are observed in tests with the two-channel implementation of the method on noisy data from a car environment. Experimental tests showed the capabilities of our two-channel implementation in terms of artifacts and SNR gain. Future work will further test the scalability of the method particularly for the three-four channel cases.

7. REFERENCES

[1] S. Gustafsson, P. Jax, A. Kamphausen, and P. Vary, "A postfilter for echo and noise reduction avoiding the problem of musical tones," in *ICASSP*, 1999, pp. 873–876.

[2] S.V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons, 2nd Edition, 2000.

[3] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[4] S.L. Gay and J. Benesty, Eds., *Acoustic Signal Processing for Teleccommunications*, Kluwer, 2001.

[5] V. Van Veen and Kevin M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, 1988.

[6] Hamid Krim and Mats Viberg, "Two decades of array signal processing research," *IEEE Signal Processing Magazine*, vol. 13, no. 4, 1996.

[7] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *ICASSP*, 1998, pp. 397–400.

[8] W. Yost, *Fundamentals of Hearing - An Introduction*, 4th Ed, Academic Press, 2000.

[9] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[10] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *Proceedings of SAM 2002, Rosslyn VA*, 2002.

[11] J. Rosca, R. Balan, N.P. Fan, C. Beaugeant, and V. Gilg, "Multichannel voice detection in adverse environments," in *Proceedings of EUSIPCO 2002*, 2002.

[12] S. Aalburg, C. Beaugeant, S. Stan, T. Fingscheidt, R. Balan, and J. Rosca, "Single- and two-channel noise reduction for robust speech recognition in car," in *ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments*, June 2002.