MULTI-CHANNEL BAYESIAN BACKGROUND NOISE SUPPRESSION USING PERCEPTUAL COST FUNCTIONS

Han Lin and Simon Godsill

Signal Processing Group
Department of Engineering, University of Cambridge, U.K
hl309@cam.ac.uk, sjg@eng.cam.ac.uk

ABSTRACT

This paper proposes a frequency-based approach for background noise suppression with consideration for human psychoacoustics. The approach utilizes a perceptual cost function analysis based on temporal masking thresholds. By optimizing the cost function, the concept eliminates background noises that mask the original signals, while maintaining the minimum perceptual distortion of the original signals. The perceptual cost function can also be implemented within a Gibbs sampling framework, which better models the uncertainty within the original signal. These approaches improve existing noise reduction techniques, enhancing perceived audio quality (PEAQ), Mean Opinion Score (MOS), and signal to noise ratio (SNR).

Index Terms— costs, noise, signal reconstruction, speech processing, statistics

1. INTRODUCTION

Using portable audio recorders or wireless communication in a noisy environment—such as recording voice mail or making news reports at a bus station—greatly reduces the perceptual quality of signals. Many noise suppression filters attempt to attenuate the background noises, sometimes with the assistance of a second microphone [10]. However, if the signal to noise ratio (SNR) is very low, suppression filters may distort the original signals while reducing the background noises, thereby, failing to improve the overall perceptual quality and intelligibility [14]. Modified filters [15] that attempt to overestimate the underlying original signals or use an extra spectral floor to mask the musical noises [2] may improve the perceptual quality, but may not necessarily optimize the perceptual quality of the signals. Therefore, this paper proposes a perceptual cost function to improve upon these filters.

2. PROBABILITY MODEL FOR THE NOISY SIGNAL

The Fourier expansion coefficients of the observed noisy signal with spectral component at frequency bin k is

$$Y_k = X_k + D_k$$

where Y_k represents the observed noisy signal, X_k represents the original signal, and D_k represents the background noise.

If we assume the Fourier expansion coefficients of the original signal X_k and the background noise D_k, they may be modelled as statistically independent, zero mean, Gaussian random variables, and the posterior probability of the magnitude of the underlying clean signal $p(a_k|Y_k)$ can be modelled as a Rician distribution [2]:

$$p(a_k|Y_k) = a_k e^{-\frac{a_k^2}{\lambda_k}} I_0(2a_k\sqrt{\frac{\upsilon_k}{\lambda_k}}) \tag{1}$$

$a_k \triangleq |X_k|$, $\lambda_k^x \triangleq E[|X_k|^2]$, $\lambda_k^d \triangleq E[|D_k|^2]$,$\frac{1}{\lambda_k} \triangleq \frac{1}{\lambda_k^x} + \frac{1}{\lambda_k^d}$, $\xi_k \triangleq \frac{\lambda_k^x}{\lambda_k^d}$, $\zeta_k \triangleq \frac{|Y_k|^2}{\lambda_k^d}$, $\upsilon_k \triangleq \frac{\xi_k}{1+\xi_k}\zeta_k$, I_n is the *Modified Bessel Function of the First Kind* of nth order.

3. THE PERCEPTUAL COST FUNCTION

We here propose a cost function $C(\hat{a}_k, a_k)$ for our estimate $\hat{a}_k$ of a_k based on the perceptual masking thresholds

$$C(\hat{a}_k, a_k) = C_1(\hat{a}_k, a_k) + w_k(|Y_k^*| - \hat{a}_k)^2 \tag{2}$$

where $C_1(\hat{a}_k, a_k)$ is a square cost function based on the masking threshold (as proposed in [16]):

$$C_1(\hat{a}_k, a_k) = \begin{cases} (\hat{a}_k - a_k)^2 - m_k^2 & \text{if } |\hat{a}_k - a_k| > m_k \\ 0 & \text{otherwise.} \end{cases}$$

- $w_k(|Y_k^*| - \hat{a}_k)^2$ is a least processing term, which is used to penalize unnecessary processing of the original audio, when the audio is active
- w_k is a weighting factor for the least processing term
- $\hat{a}_k$ is the estimated magnitude of the original signal X_k

1-4244-1484-9/08/$25.00 ©2008 IEEE

- m_k is a perceptual masking threshold
- Y_k^* is the prior estimate of original signal X_k
- γ is an approximate signal activity indicator for each frame, obtained by the threshold of the average frame energy[1]. For instance, $\gamma = 1$ for active signal frames and $\gamma = 0$ for inactive signal frames

Further, the confidence weighting w_k and the prior estimate Y_k^* are adapted as follows:

For active signal frames (i.e., $\gamma = 1$), w_k is chosen to be <0.5, and $Y_k^* = Y_k$; for inactive signal frames (i.e., $\gamma = 0$), w_k is chosen to be >3, and $Y_k^* = 0$. Generally, a_k is close to zero for inactive frames, where $\hat{a}_k$ tends to approach m_k.

The perceptual masking threshold, m_k, can be calculated by using standard procedures of both simultaneous and temporal masking, see [1] [5] [12] [11] [16] [18]

In addition, we adopt perceptual optimality criteria by using *perceptual risk* R_k concept [16] wherein R_k is defined as the expected cost:

$$\begin{aligned} R_k &\triangleq E[C(\hat{a}_k, a_k)] \\ &= \int_{-\infty}^{\infty} p_{Y_k}(Y_k) \int_{-\infty}^{\infty} C(\hat{a}_k, a_k) p_{a_k|Y_k}(a_k|Y_k) da_k d_{Y_k} \end{aligned}$$

Because our cost function $C(\hat{a}_k, a_k)$ is non-negative, to obtain the optimal Bayesian estimate, only the inner integral of the expected cost with respect to $p(a_k|Y_k)$ need be minimized. Subsequently, the optimal $\hat{a}_k$ which minimizes the *perceptual risk* R_k can be determined by the following equation:

$$\hat{a}_k^C = \arg\min_{\hat{a}_k} \underbrace{\int_{a_k} C(\hat{a}_k, a_k) p(a_k|Y_k) da_k}_{\text{inner integral of } R_k} \tag{3}$$

4. IMPLEMENTATION OF SOFT DECISION PERCEPTUAL COST FUNCTION

A soft decision hypothesis model [10] is adopted to further classify and estimate active signal components, wherein H_k^1 is used to indicate the hypothesis when audio is present, and H_k^0 is used to indicate the hypothesis when audio is not present. The binary hypothesis indicator (H_k^1/H_k^0) is applied to all signal components. When γ=0, prior probabilities are fixed as ($p(H_k^1)$=0.01 and $p(H_k^0)$=0.99).When γ=1, for each frequency bin k, prior probabilities $p(H_k^1)$ and $p(H_k^0)$ are determined from pre-classified audio signal examples of the category (male/ female) as those used in the noise suppression procedure. The posterior distribution can be modified [2]:

$$p(a_k|Y_k) = p(a_k|H_k^1, Y_k)p(H_k^1|Y_k) + p(a_k|H_k^0, Y_k)p(H_k^0|Y_k) \tag{4}$$

- $p(a_k|H_k^1, Y_k) = a_k e^{-\frac{a_k^2}{\lambda_k}} I_0(2a_k\sqrt{\frac{\upsilon_k}{\lambda_k}}) \triangleq R(a_k)$
- $p(a_k|H_k^0, Y_k) = \delta(a_k)$, $a'_k \triangleq E[R(a_k)]$
- $p(H_k^1|Y_k) \propto p(H_k^1)\frac{2|Y_k|}{\lambda_k^d} e^{-(\frac{Y_k^2 + a_k'^2}{\lambda_k^d})} I_0(\frac{2a'_k|Y_k|}{\lambda_k^d}) \triangleq U$
- $p(H_k^0|Y_k) \propto p(H_k^0)\frac{2|Y_k|}{\lambda_k^d} e^{-\frac{Y_k^2}{\lambda_k^d}} \triangleq V$

By setting the derivative of spectral amplitude estimator w.r.t. $\hat{a}_k$ to zero, the minimal risk R_k Eq. (3) can be simplified to:

$$\hat{a}_k = \frac{\bar{a}_k + w_k|Y_k^*| + \int_{\hat{a}_k - m_k}^{\hat{a}_k + m_k} (\hat{a}_k - a_k)[R(a_k)U + \delta(a_k)V] da_k}{1 + w_k} \tag{5}$$

where $\bar{a}_k$ is the mean of the posterior distribution $p(a_k|Y_k)$ Eq. (4). The integral can be solved with an iterative method:

$\hat{a}_k$ is first chosen as $\bar{a}_k$ and the integral is approximated using a *Trapezoidal Summation*; after which, the LHS of Eq. (5) can be calculated. A new $\hat{a}_k$ is then used as $\hat{a}_k$ to compute the RHS integral again. The process is iterated until $\hat{a}_k$ converges. The converged $\hat{a}_k$ value is denoted as $\hat{a}_k^{CT}$.

It remains to specify the Gaussian variable terms (λ_k^d and λ_k^x) required for the computation of Eq. (5). The first Gaussian variable λ_k^d is measured from a secondary microphone B[2]. The second Gaussian variable λ_k^x is estimated using the obtained value of λ_k^d and a standard soft-decision estimate of a_k [10], denoted as $\hat{a}_k^{ST}$.

5. IMPLEMENTATION WITH GIBBS SAMPLER

In the previous section, λ_k^x is crudely estimated using the secondary microphone B. The method could result in residual musical noise due to filtering [2]. To better estimate the uncertainty in λ_k^x, we use the Gibbs sampler[3] [3] to sample λ_k^x and a_k.

In an ideal Bayesian framework, we would wish to solve the following integral:

$$p(a_k|Y_k) = \int p(a_k|Y_k, \lambda_k^x) p(\lambda_k^x|Y_k, a_k) d\lambda_k^x$$

[1] The first few frames of the audio file is assumed to be inactive frames. The threshold value obtained by taking the average power specturm $|Y_k|^2$ of these frames is used to determine γ

[2] Microphone B is assumed to be an essentially uncontaminated measurement of the noise D_k

[3] In this version of the model, no probabilistic hypothesis switching (H_k^1/H_k^0) is applied.

$p(a_k|Y_k, \lambda_k^x)$ can be represented as a standard Rician distribution $Rice(s_k, \sigma_k)$: Eq. (1),

where $\sigma_k = \sqrt{\frac{\lambda_k^x \lambda_k^d}{2(\lambda_k^x + \lambda_k^d)}}$, $s_k = \sqrt{\frac{\lambda_k^x |Y_k|}{\lambda_k^x + \lambda_k^d}}$

In addition, because only the mean of a_k is required to minimize the perceptual cost function, $\bar{a}_k$ is determined with the mean value of Rician distribution [13]:

$$\bar{a}_k = \sigma_k \sqrt{\pi/2} L_{1/2}(-{s_k}^2/2{\sigma_k}^2)$$

where $L(x)$ denotes the Laguerre polynomial.

For active signal frames (i.e., $\gamma = 1$), $p(\lambda_k^x|Y_k, a_k)$ can be obtained using the inverted Gamma distribution [17]

$$p(\lambda_k^x|Y_k, a_k) = IG(1 + \varkappa_k, \frac{a_k^2}{2} + \psi_k)$$

A Gamma distributed prior probability is chosen [4], where $p(\lambda_k^x) = IG(\varkappa_k, \psi_k)$. $\varkappa_k$ is a constant shape parameter and ψ_k is a constant scale parameter[4].

Gibbs sampler steps are as follows, for iteration i=1,...,N

$$\begin{array}{rcl}
a_k^{(i)} & \sim & p(a_k^{(i)}|Y_k, \lambda_k^{x(i)}) = Rice(s_k^{(i)}, \sigma_k^{(i)}) \\
\bar{a}_k^{(i)} & = & \sigma_k^{(i)} \sqrt{\pi/2} L_{1/2}(-(s_k^{(i)})^2/2(\sigma_k^{(i)})^2) \\
\lambda_k^{x(i)} & \sim & p(\lambda_k^{x(i)}|Y_k, a_k^{(i)}) = IG(1+\varkappa_k, \frac{(a_k^{(i)})^2}{2}+\psi_k) \\
 & \vdots & \\
\bar{a}_k^{(N)} & = & \sigma_k^{(N)} \sqrt{\pi/2} L_{1/2}(-(s_k^{(N)})^2/2(\sigma_k^{(N)})^2) \\
\lambda_k^{x(N)} & \sim & p(\lambda_k^{x(N)}|Y_k, a_k^{(N)}) = IG(1+\varkappa_k, \frac{(a_k^{(N)})^2}{2}+\psi_k)
\end{array}$$

Finally, using the mean of the converged values of $\bar{a}_k^{(i)}$ as a new mean estimate of the original signal $\bar{a}_k$, *perceptual risk* R_k can be minimized using Eq. (5). The integral can be approximated as before and $\hat{a}_k$ can be determined by an iterative method described in the previous section. The converged $\hat{a}_k$ value is denoted as $\hat{a}_k^{MC}$. For inactive signal frames (i.e., $\gamma = 0$), a special randomizing procedure is adopted by drawing λ_k^x from the prior distribution.

6. EXPERIMENTS AND RESULTS

To determine the SNR and PEAQ scores, background noise D_k and original signal X_k are recorded separately with 8kHz sampling rate and combined subsequently. The setup attempts to simulate communication quality ($G.711$).

Two identical Electret condenser microphones are used for recordings. The two microphones (A and B) are about 30cm apart with Inter Level Difference (ILD) of about 6.5 (linear scale) which produces a small but imperceptible echo. Because of this, the ILD is not taken into account. Inter Time Difference (ITD) is assumed to be zero for the sake of simplicity.

The following examples show background noise suppression results of: an American male speaking in a noisy car park, an American male speaking next to a noisy machine fan, an American female speaking on a bus, and a British male speaking in a truck.

As shown in the tables below, the perceptual cost function improves the PEAQ scores by as much as 0.8 points, the MOS scores, by 1.7 points, and SNR, by 16.9 dB. The result is represented as $\hat{a}_k^{CT}$. When the perceptual cost function is applied to the Gibbs sampled signal, PEAQ scores, MOS scores, and SNR can be improved further by removing the residual musical noise. This signal is represented as $\hat{a}_k^{MC}$. For loud car park background noises, because of the constant shifting of phases of the noises, the PEAQ scores, MOS scores, and SNR of $\hat{a}_k^{ST}$ are worse than the observed noisy signal Y_k, as suggested by [14]. In contrast, the perceptual cost function- even with a poor estimate as mean value- is still capable of improving the PEAQ scores, MOS scores, and SNR of the observed noisy signal Y_k. Audio samples of the results of *Multi-channel Bayesian Background Noise Suppression* can be found at:

http://www-sigproc.eng.cam.ac.uk/~hl309/ICASSP2008

The paper uses the following word phrases from the *IEEE Harvard Sentences* (*Open Speech Repository)*:

I: A gold ring will please most any girl (AM)
II: When you hear the bell, come quickly (AM)
III: The beauty of the view stunned the young boy (AF)
IV: The little tales they tell are false (BM)
AM: Speech of American Male
AF: Speech of American Female
BM: Speech of British Male

Y_k, $\hat{a}_k^{ST}$, $\hat{a}_k^{CT}$, and $\hat{a}_k^{MC}$ are described in previous sections 2, 4, and 5

PEAQ scores

Audio sample	Y_k	$\hat{a}_k^{ST}$	$\hat{a}_k^{CT}$	$\hat{a}_k^{MC}$
I + car	-3.85	-3.87	-3.79	-3.78
II + fan	-3.67	-3.69	-3.18	-3.16
III + bus	-3.53	-3.02	-2.70	-2.87
IV + truck	-3.90	-3.84	-3.82	-3.80

MOS scores

Audio sample	Y_k	$\hat{a}_k^{ST}$	$\hat{a}_k^{CT}$	$\hat{a}_k^{MC}$
I + car	1.3	1.2	2.5	2.4
II + fan	1.7	2.2	3.0	3.1
III + bus	1.4	2.9	3.1	3.2
IV + truck	1.8	2.0	2.5	2.6

[4] In this implementation, consecutive signal frames (where $\gamma = 1$) are grouped together. For each group, the soft decision estimated signal $\hat{a}_k^{ST}$ is used to estimate a_k and λ_k^x of the group. These values are fitted with inverted Gamma distributions to determine the average value of $\varkappa_k$ and ψ_k

SNR (dB)

Audio sample	Y_k	$\hat{a}_k^{ST}$	$\hat{a}_k^{CT}$	$\hat{a}_k^{MC}$
I + car	4.01	2.34	4.77	4.80
II + fan	2.49	2.86	5.10	5.84
III + bus	-14.11	2.35	2.79	2.92
IV + truck	-14.42	-1.00	0.47	0.48

7. CONCLUSION

The examples in the previous sections show the perceptual cost function can improve frequency domain noise suppression filters; the examples also show how the perceptual cost function can be implemented within a simple Gibbs sampling framework. More complex sampling methods using Markov Chain Monte Carlo (MCMC) [4] [3] [17] can be adopted to improve the estimates of the original signal.

The perceptual cost function can also be applied to time domain noise suppression filters in a time frequency framework, where the time domain filtered signal is transformed into frequency domain and then applied with perceptual cost function suppression. Examples are GSM Buzz (Electromagnetic Interference) removal [9] and missing audio interpolation (click removal) [4] [8].

In addition, if two microphones are placed very close to each other, ILD cues can be used to produce more accurate results. An accurate filter could be adapted by minimizing cost functions of ILD cues [7].

8. REFERENCES

[1] A. Chen, N. Shehad, A. Virani, Erik Welsh, W.A.V.S Compression, //is.rice.edu/~welsh/elec431/index.html

[2] Y. Ephraim and D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-32(6):1109-1121, 1984

[3] W.R. Gilks, S. Richardson, D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, London, Chapman and Hall, 1996

[4] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration - a statistical model-based approach*, Springer-Verlag, 1998

[5] T. S. Gunawan and E. Ambikairajah, On the Use of Simultaneous and Temporal Masking for Noise Suppression in Cochlear Implant, Eleventh Australasian International Conference on Speech Science and Technology, 2006

[6] S. Kim, J. Lee, and D.Sung, A Shifted Gamma Distribution Model for Long-Range Dependent Internet Traffic, IEEE Communications Letters, Vol. 7, No. 3, March 2003

[7] T. J. Klasen, S. Doclo, T. V. Bogaert, M. Moonen, J. Wouters, Binaural multi-channel Wiener filtering for hearing aids: Preserving interaural time and level differences, IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2006

[8] H. Lin and S. J. Godsill, The Multi-channel AR Model For Real-time Audio Restoration. IEEE Workshop on Audio and Acoustics, Mohonk, NY State, October 2005

[9] H. Lin and S. J. Godsill, Real-Time Bayesian GSM Buzz Removal, Proc. of the 9th Int. Conference on Digital Audio Effects, Montreal, Canada, September 2006

[10] R. J. McAulay and M. L. Malpass, Speech enhancement using a soft-decision noise suppression filter. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-34(4):744-754, 1986

[11] W. Pai, Method for utilizing temporal masking in digital audio coding, US Patent 6895,374 B1, May 17th, 2005

[12] T. Painter, A. Spanias; Perceptual coding of digital audio, Proceedings of the IEEE Volume 88, Issue 4, April 2000 Page(s):451 - 515

[13] S. O. Rice, Statistical properties of a sine wave plus random noise. Bell System Technical Journal, 27:109-157, 1948

[14] R. C. Rohlfs, Speech enhancement using the Widrow-Hoff adaptive noise-canceling tapped delay-line filter; A CSP30 implementation," MITRE Tech. Rep. MTR-3626, Mar, 1979

[15] D. E. Tsoukalas, J. Mourjopoulos, G. Kokkinakis, Perceptual filters for audio signal enhancement. Journal of the audio Engineering Society, 45, 22-36, 1997

[16] P. J. Wolfe and S. J. Godsill, Perceptually Motivated Approaches to Music Restoration, Journal of New Music Research, Vol. 30, No. 1, pp. 83-92, 2001

[17] P. J. Wolfe, S. J. Godsill, and W. J. Ng, Bayesian variable selection and regularization for time-frequency surface estimation, J. Roy. Stat. Soc. Ser. B, vol. 66, pp. 575–590, 2004

[18] E. Zwicker, Subdivision of the audible frequency range into critical bands (Frequencz-gruppen). Journal of the Acoustic Society of America, 33(2):248, 1961