2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

심리음향모델에 근거한 음성개선

이진걸
배재대학교 전자공학과

Speech Enhancement Based on Psychoacoustic Model

Jingeol Lee
Department of Electronic Engineering, Paichai University

E-mail: jingeol@mail.paichai.ac.kr

Abstract - The perceptual filter for speech enhancement was analytically derived where the frequency content of the input noisy signal was made the same as that of the estimated clean signal in auditory domain. However, the analytical derivation should rely on the deconvolution associated with the spreading function in the psychoacoustic model, which results in an ill-conditioned problem. In order to cope with the problem associated with the deconvolution, we propose a novel psychoacoustic model based speech enhancement filter whose principle is the same as the perceptual filter, however the filter is derived by a constrained optimization which provides solutions to the ill-conditioned problem.

Index Terms - Speech Enhancement, Psychoacoustic Model, Masking, Optimization

I. INTRODUCTION

Speech enhancements have been actively studied for facilitating human-machine interface and mobile communications in noisy environments. Recently, speech enhancement methods incorporating psychoacoustic model, which is already widely used in perceptual wideband audio coding, have been introduced. One of them is the perceptual filter, which is analytically derived where the frequency content of input noisy signal is made the same as that of the estimated clean signal in auditory domain [1]. However, we note that the analytical derivation of the perceptual filter is possible by assuming that the filter remains the same for all time frames and critical bands, leading to the oversimplified filter formula. The perceptual filter should rely on the deconvolution associated with the spreading function in the psychoacoustic model, which results in an ill-conditioned problem [2]. In order to cope with the problem associated with the deconvolution, we propose a novel psychoacoustic model based speech enhancement filter whose principle is the same as the perceptual filter, however the filter is derived by a constrained optimization which provides a solution to the ill-conditioned problem.

II. PSYCHOACOUSTIC MODEL BASED SPEECH ENHANCEMENT FILTER

A linear filter $H(b,i)$ is introduced whose gain is assumed to be constant within the same critical band. The power spectrum of the enhanced speech signal is given by

$$\hat{X}_p(k,i) = H(b,i)Y_p(k,i), \quad k_{lb} \le k \le k_{hb}, \quad 0 \le b \le B-1 \qquad (1)$$

where $Y_p(k,i)$ is the power spectrum of the noisy speech signal at the analysis frame index i, and b is the critical band index, and k_{lb} and k_{hb} are the lower and upper bounds, respectively, of the critical band b, and B is the total number of critical bands. Considering that the psychoacoustic representation of the noisy speech signal at a certain frequency is found by summing the spreaded powers in adjacent critical bands, the psychoacoustic representation at that frequency can be modified by weighting the powers in adjacent critical bands as

$$\sum_{v=0}^{B-1} \{SS[v,b(j)]H(v,i)Y_t(v,i)\} = X_f[b(j),i], \quad 0 \le b \le B-1 \qquad (2)$$

where j is a frequency index, and $Y_t(b,i)$ is the time-domain smeared power, and SS is the spreading function, and X_f is the psychoacoustic representation of the clean speech signal. As shown in (2), $SS[v,b(j)]Y_t(v,i)$ is the spreaded power at frequency index j corresponding to the critical band index of $b(j)$ from the power of $Y_t(v,i)$ at the critical band index v. Evaluation of (2) results in the linear algebraic equation in the form $Y \cdot \boldsymbol{h} = \boldsymbol{x}$. Y is a matrix whose size is the number of frequencies evaluated by the number of critical bands, and whose elements of each row are the spreaded powers at the frequency of evaluation from the power of $Y_t(v,i)$ at the corresponding critical band. The vector $\boldsymbol{h}$ consists of the filter coefficients $H(b,i)$, whose size is exactly the number of critical bands. The vector $\boldsymbol{x}$ consists of the psychoacoustic representation X_f at the frequencies of evaluations, whose size is exactly the number of frequencies evaluated. The filter coefficients can be solved by the method based on singular value decomposition (SVD). However, it is found that the SVD based solution occasionally results in negative values as the

filter coefficients depending on noise level, which gives rise to negative powers. In order to cope with this problem, the problem is formulated as a constrained optimization problem as follows:

$$\min_{h} \|Y \cdot h - x\|_2 \quad such\ that \quad 0 \le h_0,\ h_1,\ \ldots,\ h_{B-1} \le 1 \qquad (3)$$

III. EXPERIMENTAL RESULTS

The perceptual filter and the proposed speech enhancement filter are tested for comparison with artificially generated signals of the sampling rate of 8 KHz. In the test, the psychoacoustic model, originally developed for MPEG audio coding, is modified to accommodate the test signals [3]. The psychoacoustic representation of signals can be obtained by removing the masking index and the absolute threshold terms from the masking threshold in the MPEG audio [3], [4].

A sinusoidal signal with the amplitude of 1000 and the frequency of 1000 Hz and a pseudorandom noise in the range of -50 to 50 are generated for demonstrating the validity of the proposed speech enhancement filter in contrast to the perceptual filter. The psychoacoustic representations of the noisy signal produced by adding the sinusoidal signal and the noise, the sinusoidal signal, and the enhanced signal are shown in Fig. 1(a) and (b) for the perceptual filter and the proposed filter, respectively. It is shown in Fig. 1 that the perceptual filter produces the psychoacoustic representation of the enhanced signal different from that of the sinusoidal signal whereas the proposed filter provides exact solution except in the narrow frequency region around 600 Hz. In contrast to the perceptual filter, the proposed filter seeks a solution taking spreadings of adjacent critical bands into account, and thus the psychoacoustic representation of the enhanced signal results in almost a perfect rendition of the sinusoidal signal. The negligible discrepancy from the psychoacoustic representation of the sinusoidal signal around 600 Hz is attributed to the numerical error associated with the optimization.

IV. CONCLUSIONS

We propose a novel psychoacoustic model based speech enhancement filter derived by formulating the problem as a constrained optimization, by which the frequency content of the input noisy signal is made the same as that of the estimated clean signal in auditory domain as the analytically derived perceptual filter. It is demonstrated with a sinusoidal signal and random noise that the proposed filter produces exact solutions in contrast to the perceptual filter.

ACKNOWLEDGEMENT

The author wishes to acknowledge the financial support of the Korea Research Foundation made in the program year of 1998.

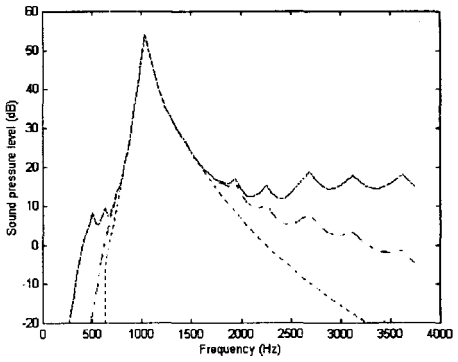
REFERENCES

[1] Dionysis E. Tsoukalas, John Mourjopoulos, and George Kokkinakis, "Perceptual Filters for Audio Signal Enhancement," J. Audio Eng. Soc., Vol. **45**, No. 1/2, pp. 22-36, Jan./Feb. 1997

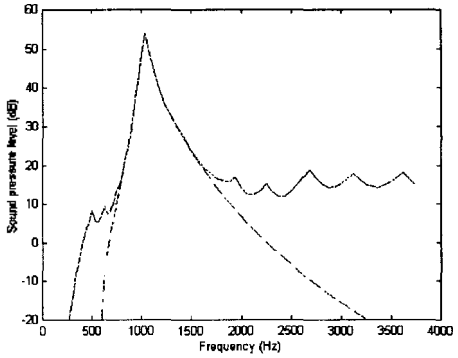
[2] Raymond N. J. Veldhuis, "Bit Rates in Audio Source Coding," IEEE Journal on Selected Areas in Communications, Vol. **10**, No. 1, pp. 86-96, Jan. 1992

[3] ISO/IEC 11172-3, "Information technology-Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s Part3: Audio"

[4] Eberhard Zwicker and U. Tilmann Zwicker, "Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory System," J. Audio Eng. Soc., Vol. **39**, No. 3, pp. 115-126, Mar. 1991



(a)



(b)

Fig. 1. Psychoacoustic representations

(a) Perceptual filter (b) Proposed filter
—— Noisy signal
-·-·- Enhanced signal
····· Sinusoidal signal