Speech Enhancement Based on Psychoacoustic Model

*Jingeol Lee, **Soowon Kim

Abstract

Psychoacoustic model based methods have recently been introduced in order to enhance speech signals corrupted by ambient noise. In particular, the perceptual filter is analytically derived where the frequency content of the input noisy signal is 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. It is demonstrated with artificially generated signals that the proposed filter operates according to the principle. It is shown that superior performance results from the proposed filter over the perceptual filter provided that a clean speech signal is separable from noise. (Classification No. 3.2)

I. Introduction

Speech enhancements have been actively studied for facilitating human-machine interface and mobile communications in noisy environments[1]-[8]. The short-time spectral amplitude (STSA) -based methods including the spectral subtraction rely on the assumption that speech signals and noise are uncorrelated. Also, the STSA-based methods take advantage of the unimportance of the short-time phase. Therefore, these methods modify the amplitude while preserving the phase of input noisy speech signal for the reconstruction of the estimated clean speech signal.

Recently, speech enhancement methods incorporating psychoacoustic model, which is already widely used in perceptual wideband audio coding, have been introduced [4], [6]-[8]. The psychoacoustic model is a mathematical model of the masking behavior of the human auditory system. The masking is a perceptual property, by which the presence of a strong signal makes the spectral and temporal neighborhood of weaker signals imperceptible [9]. Empirical results also show that the human auditory system has a limited, frequency-dependent resolution over which the human ear seems to integrate. This dependency can be expressed in terms of critical bandwidth of 100Hz for frequencies below 500Hz and approximately a third octave for frequencies above 500Hz [9], [10]. The speech enhancement methods incorporating the psychoacoustic model can be classified as the STSA-based methods since the short-time amplitude is modified; however, the modification takes the auditory system into account. 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. On the other hand, most psychoacoustic model based methods exploit the masking property of the auditory system. The input noisy signal is processed with linear or nonlinear filtering such that the audible parts of the noise are masked by the estimated clean signal or in the least, the best tradeoff between noise reduction and speech quality is made. However, we note that the perceptual filter should rely on the deconvolution associated with the spreading function in the psychoacoustic model, which results in an ill-conditioned problem[10], [11]. In addition, we find that masking property based methods cannot mask audible parts of the noise completely because the methods process the noise such that their psychoacoustic representations fall below the masking thresholds of the estimated clean speech signal, which results in reducing the masking threshol at the same time for a single-microphone situation. Therefore, the psychoacoustic representation of the noise remains above the masking threshold unless the psychoacoustic representation is reduced to absolute thresholds of hearing, which may cause drastic distortions of the speech signal.

In order to cope with aforementioned problems, 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

* Department of Electronic Engineering Paichai University
** Department of Electronic Engineering, Korea University

Manuscript Received : May 29, 2000

ill-conditioned problem. It is demonstrated with a sinusoidal signal and a random noise that the proposed filter operates according to the principle in contrast to the perceptual filter. It is shown that superior performance results from the proposed filter than the perceptual filter assuming separable clean speech signal from noise.

The perceptual filter is reviewed in Section II. In Section III, the novel psychoacoustic model based speech enhancement filter utilizing the constrained optimization is proposed. In Section IV, its performances are demonstrated by comparing experiment results of the perceptual filter and the proposed filter. The conclusions are drawn in Section V.

II. Reviews of the perceptual filter

The perceptual filter exploits the psychoacoustic representations of signals that include the time and the frequency-domain smearing in the auditory system[4]. Let $x(n)$ be a discrete-time signal. This signal is transformed to the frequency-domain representation according to overlap addition method,

$$X_w(k,i)=\sum_{n=0}^{N-1} w(n)x(n+off_i)e^{-j2\pi nk/N} \qquad (1)$$

where $w(n)$ is a window function, and N is the Fourier transform length, and off_i is the window drifting factor. The power spectrum of the signal is given by

$$X_p(k,i)=|X_w(k,i)|^2,\quad 0\le k\le N-1. \qquad (2)$$

From the power spectrum, the total energy per critical band is calculated as

$$X_a(b,i)=a_0(b)\sum_{k=k_{lb}}^{k_{hb}} X_p(k,i),\quad 0\le b\le B-1 \qquad (3)$$

where b is the critical band index, and $a_0(b)$ is an outer-to-inner ear transformation function, 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. Then, the time-domain smearing is described by

$$X_t(b,i)=X_a(b,i)+T_f(b)X_t(b,i-1),\quad 0\le b\le B-1 \; (4)$$

where $T_f(b)$ is an exponential function. The function $X_t(b,i)$ is then convolved with the basilar membrane spreading function, which provides the frequency-domain smearing.

$$X_f(b,i)=\left\{\sum_{v=0}^{b-1}\left[S_2(v,b-v)[X_t(v,i)]^{1+0.02(b-v)dz}\right]^{\delta/2}+\sum_{v=b}^{B-1}[S_1(v-b)X_t(v,i)]^{\delta/2}\right\}^{2/\delta}$$

$$,\ 0\le b\le B-1 \qquad (5)$$

As shown in Eq. (5), the spreading function is expressed by two different functions, S_1 for frequencies above the critical band b, and S_2 for frequencies below the band b. A linear filter $H(b,i)$ is introduced whose gain is assumed to be constant within the same critical band. The enhanced speech signal is given by

$$\hat{X}_p(k,i)=H(b,i)Y_p(k,i),\ k_{lb}\le k\le k_{hb},\quad 0\le b\le B-1 \quad (6)$$

where $Y_p(k,i)$ is the power spectrum of the noisy speech signal. The perceptual filter modifies the power spectrum of the noisy speech signal so that the resulting psychoacoustic representation is the same as that of the clean speech signal.

$$\hat{X}_f(b,i)=X_f(b,i),\quad 0\le b\le B-1 \qquad (7)$$

With the assumptions of $\delta=2$, $1+0.002(b-v)dz\approx 1$ and with some mathematical manipulations, Eq. (7) becomes

$$\sum_{v=0}^{B-1}\left\{SS(v,b)a_0(v)\sum_{m=0}^{i}[T_f^{i-m}(v)H(v,m)Y_M(v,m)]\right\}=X_f(b,i)$$

$$,0\le b\le B-1 \qquad (8)$$

where the spreading function SS includes S_1 and S_2 in Eq. (5), and $Y_M(v,m)=\sum_{k=k_{lb}}^{k_{hb}} Y_p(k,m)$. The external summation in Eq. (8) corresponds to the frequency-domain smearing whereas the nested summation corresponds to the time-domain smearing. It is assumed that the enhancement process is performed by the same filter $H(b,i)$ for all time frames and critical bands.

$$H(v,m)=H(b,i),\quad 0\le v\le B-1,\quad 0\le m\le i \qquad (9)$$

Substutution of Eq. (9) into Eq. (8) becomes

$$H(b,i)\sum_{v=0}^{B-1}\left\{SS(v,b)a_0(v)\sum_{m=0}^{i}[T_f^{i-m}(v)Y_M(v,m)]\right\}=X_f(b,i)$$

$$,\ 0\le b\le B-1. \qquad (10)$$

The summation on the left-hand of Eq. (10) is the psychoacoustic representation of the noisy speech signal. Therefore, the time-frequency model dependent filter can finally be expressed as

$$H_{tf}(b,i)=\frac{X_f(b,i)}{Y_f(b,i)},\ 0\leq b\leq B-1. \qquad (11)$$

The perceptual filter that considers only the frequency-domain smearing is given by

$$H_f(b,i)=\frac{X'_f(b,i)}{Y'_f(b,i)},\ 0\leq b\leq B-1 \qquad (12)$$

where $X'_f(b,i)=X_f(b,i)\mid_{T_f(b)=0}$.

The analytical derivation of the perceptual filter is possible by assuming that the filter remains the same for all time frames and critical bands as in Eq. (9). This assumption is taken in consideration of the fact that the psychoacoustic representation is a very slowly varying function with the time and the frequency-domain smearing. However, the filter gains result in independent from those of adjacent critical bands, which have to be interrelated due to the spreading function. Therefore, the previous assumptions have led to the oversimplified psychoacoustic representations of signals. The filter gains may be similar in adjacent critical bands and time frames, however it is not appropriate to assume that the gains are the same for all time frames and critical bands.

III. Psychoacoustic Model Based Speech Enhancement Filter

The analytical derivation of an enhancement filter involves the deconvolution associated with the spreading function in the psychoacoustic model as in Eq. (8), which results in an ill-conditioned problem. The approach often leads to artifacts such as negative energy for the estimated speech signal [10], [11]. In order to cope with problems associated with the deconvolution, we propose a novel psychoacoustic model based speech enhancement filter whose principle is the same as the perceptual filter, however derived by a constrained optimization.

Since powers of the spectral lines are summed within each critical band to form the psychoacoustic representation of the noisy speech signal $Y_f(b,i)$, the filter gain $H(b,i)$ is assumed to be constant within each critical band as shown in Eq. (6).
Considering that the psychoacoustic representation of the noisy speech signal at a certain frequency is found by summing the spreaded powers of $Y_t(b,i)$ in adjacent critical bands, the psychoacoustic representation at that frequency can be modified by weighting the powers of $Y_t(b,i)$ in adjacent critical bands. Therefore, Eq. (8) can be expressed as

$$\sum_{v=0}^{B-1}\{SS[v,b(j)]H(v,i)Y_t(v,i)\}=X_f[b(j),i],\ 0\leq b\leq B-1 \qquad (13)$$

where j is a frequency index. As shown in Eq. (13), $SS[v,b(j)]Y_t(v,i)$ is the spreaded power of $Y_t(b,i)$ 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 Eq. (13) at the properly chosen frequency set 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 of $Y_t(b,i)$ 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 $X_f(b,i)$ at the frequencies of evaluations, whose size is exactly the number of frequencies evaluated. The number of unknown, i.e., filter coefficients can be greater or less than the number of the equations depending on the frequency set, which 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:

$$\min_{h}\ \| Y\cdot \boldsymbol{h}-\boldsymbol{x}\|_2\ \textit{such that}\ 0\leq h_0,h_1,\ldots,h_{B-1}\leq 1 \qquad (14)$$

It is expected that the resulting psychoacoustic representation $\hat{X}_f(b,i)$ may be somewhat different from the constrained psychoacoustic representation $X_f(b,i)$ because the spreading function depends on the power level of $Y_t(b,i)$. Since the higher power of $Y_t(b,i)$ is spreaded more gradually, the resulting psychoacoustic representation of the estimated clean speech signal $\hat{X}_f(b,i)$ is expected to fall somewhat below the constrained psychoacoustic representation $X_f(b,i)$. However, it is found from a series of experiments as will be shown in the next section that the psychoacoustic representation of the estimated clean signal results in closer to the clean signal than the perceptual filter.

IV. Experimental results

The perceptual filter and the proposed speech enhancement filter are tested for comparison with both artificially generated signals and real speech signals. In both tests, the psychoacoustic model, originally developed for MPEG audio coding, is modified to accommodate the test signals[12]. The psychoacoustic model 1 of MPEG audio supporting the sampling rate of 32KHz and the frame size of 1024 samples is modified to accommodate the test signals with the sampling rate of 8KHz. Accordingly, the frame size is reduced from 1024 samples to 256 samples to maintain the same frequency resolution of the FFT. The number of subsampled frequencies, at which the masking thresholds are evaluated in the MPEG audio, is reduced from 132 to 78, and the number of the critical bands is reduced from 24 to 17, which covers baseband of the test signals, 0 - 4kHz. 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 [12], [13]. Since the psychoacoustic model considers only the frequency-domain smearing, the perceptual filter, which includes only the frequency-domain smearing, is implemented for the comparison tests.

A sinusoidal signal with the amplitude of 1000 and the frequency of 1000Hz 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. The psychoacoustic representations of the signals are evaluated at the subsampled frequencies that span critical bands. Therefore, the gain of the perceptual filter is determined by taking the ratio of the maximum value of the psychoacoustic representation of the sinusoidal signal and the maximum value of the noisy signal in each critical band as in Eq. (12). On the other hand, the proposed filter processes the noisy signal according to the Eq. (14) using the psychoacoustic representation of the sinusoidal and the noisy signal, evaluated at the 78 subsampled frequencies. The initial values of elements of h are assigned to 1s, which corresponds to no modification of the noisy signal. The psychoacoustic representations of the enhanced signals are calculated using the frequency contents of the enhanced signals using Eq. (6).

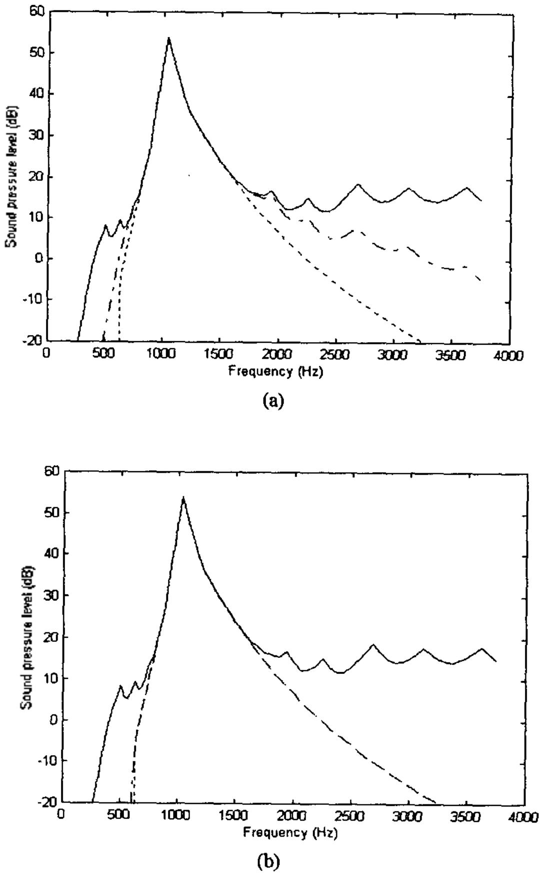
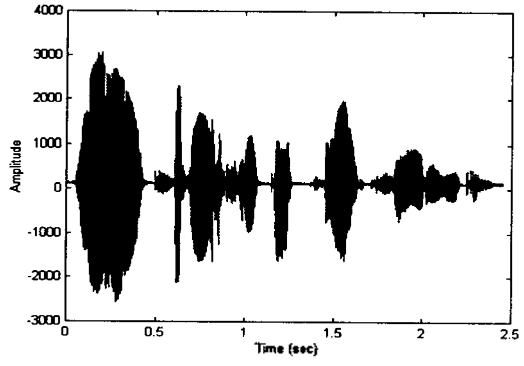


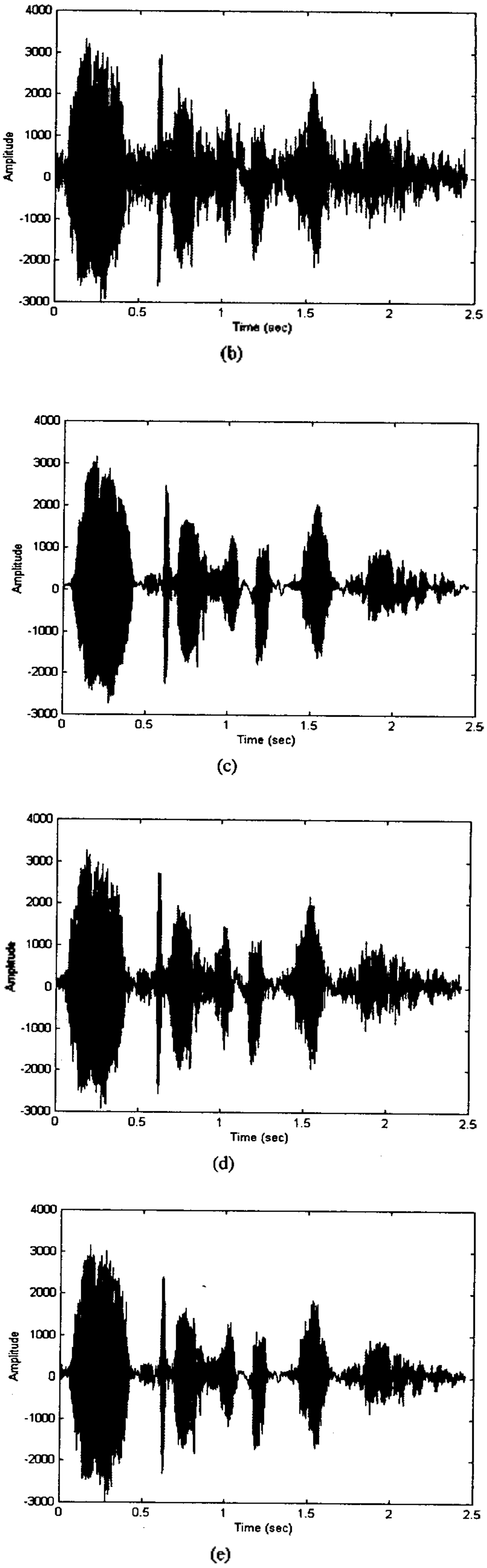
Figure 1. Psychoacoustic representations,
(a) Perceptual filter, (b) Proposed filter.
—— Noisy signal — · — · — · Enhanced signal
.... Sinusoidal signal

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 600Hz. This is caused by the fact that the perceptual filter merely adjusts peak values of the psychoacoustic representation of the noisy signal at each critical band such that the peak values become the same as the corresponding psychoacoustic representations of the sinusoidal signal. Therefore, the psychoacoustic representation of the enhanced signal results in higher than that of the sinusoidal signal due to the spreading of the sinusoidal and the noise components, which is prominent in noise dominant regions in Fig. 1 (a). Moreover, it is expected that the processed noise becomes audible in the perceptual filter considering that the psychoacoustic representations are higher than the masking thresholds by the masking index [13]. On the other hand, 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 600Hz is attributed to the numerical error associated with the optimization.

The above test is performed to demonstrate the validity of the proposed speech enhancement filter using a single frame of the artificial signals. As a second test, the enhancement of the bus noise-corrupted female speech signal by the theoretical STSA limit is compared against both the perceptual filter and the proposed filter assuming that the clean speech signal and the noise are separable. We adopt this assumption in order to rule out the signal distortions caused by the estimation of the noise in single-microphone situation. The theoretical STSA limit is obtained by reconstructing the speech signal using the spectral amplitudes of the clean signal combined with the phases of the noisy signal while adjusting the amplitude of the noise such that the distortions are not perceived in the reconstructed speech signal. Therefore, the STSA limit, as its name implies, is theoretically the best obtainable enhanced speech signal for the STSA-based method. The amount of noise allowed by the STSA limit is the maximum perceptually suppressible noise level. We perform this test under such worst condition in order to make the comparison distinctive. In this test, FFT size is set equal to the frame size of 256, and it was shown that the temporal aliasing caused by circular convolution is negligible [2]. The analysis frames are overlapped with adjacent frames by 50% and the enhanced speech signal is obtained by the overlap addition method.



(a)



(b)

(c)

(d)

(e)

Figure 2. Time-domain plots,
(a) Clean speech, (b) Noisy speech, (c)STSA limit, (d) Perceptual filter, (e) Proposed filtet.

Time-domain plots for the clean speech signal, the noisy speech signal, the enhanced speech signals by the STSA limit, by the perceptual filter, and by the proposed filter are shown in Fig. 2. The enhanced speech signal by the proposed filter is closer to the STSA limit, whereas the perceptual filter produces relatively a noisier speech signal, which is consistent with the result of the first test with the artificial signals. For numerical comparisons, the following objective evaluation is performed using the *SNR* measurement, defined as

$$SNR_i = 10 \log_{10} \frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} d(n)^2} (dB)$$

$$SNR_o = 10 \, log_{10} \frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} [\varnothing(n) - x(n)]^2} (dB) \qquad (15)$$

where $x(n)$ is the clean speech signal, $d(n)$ is the additive noise, $\varnothing(n)$ is the signal under the measurement, i.e., the enhanced speech signal, and N is the length of the signals. The SNR_i and the SNR_o are the *SNR* of the noisy speech signal and the enhanced speech signal, respectively, and the difference between two values indicates the *SNR* improvement through the enhancements. The *SNR* measurements for the female speech sentence in Fig. 2 and an additional male speech sentence with the bus noise allowed by the STSA limit and with half of the noise corresponding to the STSA limit are depicted in Table 1. Along with that, the pair comparison results for subjective speech quality assessment are given. In the pair comparisons, listeners are played each pair twice and asked to choose the version they prefer.

Table 1. *SNR* measurements and pair comparisons.

Female Speech Sentence				
$SNR_i(dB)$	Method	$SNR_0(dB)$	Preference(%)	Not Sure(%)
7.8	Perceptual	10.8	0.0	0.0
(STSA Limit)	Proposed	12.2	100.0	
13.9	Perceptual	15.7	5.0	5.0
	Proposed	16.0	90.0	

Male Speech Sentence				
$SNR_i(dB)$	Method	$SNR_0(dB)$	Preference(%)	Not Sure(%)
5.9	Perceptual	8.5	9.5	9.5
(STSA Limit)	Proposed	9.5	81.0	
11.9	Perceptual	13.6	5.0	0.0
	Proposed	14.1	95.0	

Table 1 shows that the proposed filter outperforms the perceptual filter in the *SNR* measurements, and with the pair comparisons, the enhanced speech signal by the proposed filter sounds closer to the clean speech signal than that by the perceptual filter. In addition, the pair comparisons reveal that the perceptual filter produces residual bus noise in the enhanced speech signal.

V. Conclusions

We propose a novel psychoacoustic model based speech enhancement filter, 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 perceptual filter. The perceptual filter is analytically derived by assuming that the filter remains the same for all time frames and critical bands, leading to the oversimplified psychoacoustic representations of signals. 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, the proposed filter is derived by formulating the problem as a constrained optimization.

It is demonstrated with a sinusoidal signal and random noise that the proposed filter produces exact solutions whereas the perceptual filter produces a psychoacoutic representation of the enhanced signal, different from that of the sinusoidal signal. For the speech signal corrupted by the STSA limit, it is shown that the enhanced speech signal by the proposed filter is closer to the STSA limit, whereas the perceptual filter produces relatively a noisier speech signal. In addition, the SNR measurements for objective speech quality measure and the pair comparisons for subjective speech quality measure support the superiority of the proposed filter over the perceptual filter.

Acknowledgements

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

References

1. Jae S. Lim and Alan V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," Proceedings of the IEEE, Vol. 67, No. 12, pp. 1586-1604, Dec. 1979.
2. Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-27, No. 2, pp. 113-120, Apr. 1979.

3. Douglas OShaughnessy, "Enhancing Speech Degraded by Additive Noise or Interfering Speakers," IEEE Communications Magazine, pp. 46-52, Feb. 1989.
4. 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.
5. Robert. J. McAulay and Marilyn. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-28, No. 2, pp. 137-145, Apr. 1980.
6. Dionysis E. Tsoukalas, John N. Mourjopoulos, and George Kokkinakis, "Speech Enhancement Based on Audible Noise Suppression," IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 6, pp. 497-514, Nov. 1997.
7. Nathalie Virag, "Speech Enhancement Based on Masking Properties of the Auditory System," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 796-799, 1995.
8. Stefan Gustafsson, Peter Jax, and Peter Vary, "A Novel Psychoacoustically motivated Audio Enhancement Algorithm Preserving Background Noise Characteristics," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 397-400, May 1998.
9. Davis Pan, "A Tutorial on MPEG/ Audio Compression," IEEE Multimedia Magazine, Vol. 2, No. 2, pp. 60-74, Jun. 1995.
10. 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.
11. James D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," IEEE Journal on Selected Areas in Communications, Vol. 6, No. 2, pp. 314-323, Feb. 1988.
12. 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"
13. 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.

▲ Jingeol Lee

Jingeol Lee received the Bachelor of Engineering degree; the Master of Science degree from Korea University in 1981 and 1985, respectively; and the Doctor of Philosophy degree from University of Florida in 1994. From 1982 to 1990, he was with Agency for Defense Development, and from 1995 to 1996, he had worked for Samsung Electronics, where he was involved in research and development of military electronics throughout his job experiences. Since 1997 he has been with the Department of Electronic Engineering, Paichai University, where he is engaged in researches in voice CODEC and remote sensing.

▲ Soowon Kim

Soowon Kim received the B.S. degree from Korea University in 1974, M.S. and Ph.D. degree from Texas A&M University in 1983 and 1987, respectively. In 1987 he joined the faculty of the Department of Electronics and Engineering at Korea University, Korea. He was section head of the Circuit Design Section of the Inter-university Semiconductor research Center from 1987 to 1989. Since 1992 he is working as consulting staff member for the Ministry of Information and Communication. Since 1994 he is a consulting staff member for the Ministry of Science and Technology and Commerce. His professional interests include ASIC and VLSI design, circuit design, special purpose processor architecture, memory structure and hierarchy.