# Warped Discrete Cosine Transform-Based Noisy Speech Enhancement

Joon-Hyuk Chang, *Member, IEEE*

***Abstract*—In this paper, a warped discrete cosine transform (WDCT)-based approach to enhance the degraded speech under background noise environments is proposed. For developing an effective expression of the frequency characteristics of the input speech, the variable frequency warping filter is applied to the conventional discrete cosine transform (DCT). The frequency warping control parameter is adjusted according to the analysis of spectral distribution in each frame. For a more accurate analysis of spectral characteristics, the split-band approach in which the global soft decision for speech presence is performed in each band separately is employed. A number of subjective and objective tests show that the WDCT-based enhancement method yields better performance than the conventional DCT-based algorithm.**

***Index Terms*—Discrete cosine transform (DCT), frequency warping control parameter, Laguerre filter, speech enhancement, split-band global soft decision, warped DCT (WDCT).**

## I. Introduction

RECENTLY, there has been an increasing interest in noisy speech enhancement for speech coding and recognition since the presence of noise seriously degrades the performance of the systems. Many approaches have been investigated in order to achieve speech enhancement. These include the spectral subtraction, Wiener filtering, soft decision estimation, and minimum mean-square error (MMSE) estimation approaches [1]–[4]. Most of this research on speech enhancement is based on the discrete Fourier transform (DFT) to make it easier to eliminate the noise from the noisy speech in the frequency domain. However, the discrete cosine transform (DCT) has been found to be better at enhancing noisy speech as compared to the DFT because of several advantages [5]. The main reason is that the DCT provides a significantly higher energy compaction capability compared to the DFT. To provide a higher resolution for the energy compacted region without increasing the DCT length, a method to warp the input frequency is devised to adjust the frequency distribution of the input speech to be more suitable for the DCT. Specifically, an online approach to speech enhancement for the warped DCT (WDCT) is proposed. Moreover, the incorporation of the global soft decision for split-bands leads to a robust determination of the frequency warping control parameter. It will be shown in this paper that the WDCT outperforms the conventional DCT in the mean opinion score (MOS) and segmental signal-to-noise ratio (SNR) tests for the enhancement of noisy speech. Furthermore, our approach can be implemented in real time with a little additional computational complexity. This basic idea was originally proposed in [6], and this paper gives an in-depth description as well as extensive experimental results.

The organization of this paper is as follows. A definition and implementation of the WDCT are given in Section II. The speech enhancement algorithm for the WDCT and the frequency warping control parameter determination are described in Section III. In Section IV, a number of subjective and objective quality tests are conducted to evaluate the performance, and, finally, in Section V, some concluding remarks are drawn.

Manuscript received August 26, 2004; revised November 4, 2004 and January 14, 2005. This work was supported by the Post-Doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF). This paper was recommended by Associate Editor H. Leung.

The author was with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA. He is now with the Department of Electronic Engineering, Inha University, Incheon 402-751, Korea (e-mail: changjh@hi.snu.ac.kr).

Digital Object Identifier 10.1109/TCSII.2005.850448

## II. Warped Discrete Cosine Transform

### A. Review

The $M$-point DCT $\{Y_0, Y_1, \cdots, Y_{M-1}\}$ of a length-$M$ input sequence $y[n]$, $0 \le n \le M-1$, is defined by

$$Y_k = U(k) \sum_{n=0}^{M-1} y_n \cos \frac{(2n+1)k}{2M}\pi, \quad \text{for} \quad 0 \le k \le M-1 \tag{1}$$

where

$$U(k) = \begin{cases} \frac{1}{\sqrt{2}}, & k = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

The $k$th row of the $M \times M$ DCT matrix can be viewed as a filter whose transfer function is given by

$$F_k(z) = \sum_{n=0}^{M-1} U(k) \cos \frac{(2n+1)k\pi}{2M} z^{-n}. \tag{3}$$

That is, the $n$th coefficient of $F_k(z)$ is the $(k,n)$th element of the DCT matrix. It can be shown that $F_k(z)$ is a bandpass filter with a center frequency at $(2k+1)/2M$ when the sampling frequency is normalized to 1. Hence, the magnitude of the output of $F_k(z)$ for small $k$ is generally larger for low-frequency inputs such as voiced sounds, which enables data compression by giving more emphasis to the lower band outputs than the higher band ones [7]. On the other hand, for an input with mostly high-frequency components, the magnitude of the output from $F_k(z)$ for higher $k$ is large. This is a desirable feature for noise-removal purposes [5].

There are a few factors affecting the discrimination of speech. In particular, frequency selectivity is one of the important aspects in speech discrimination [8]. Definition of the frequency selectivity is that, with intensive care, listeners want to hear

important frequency regions which are mainly high spectral magnitude areas. For this reason, providing higher resolution in a selected frequency range like a high spectral magnitude region is necessary [9]. A possible solution is to increase the total DCT size which depends on the distance between the two consecutive sampling points in the frequency domain. This increases the frequency resolution and improves the speech quality. However, it will increase the computational complexity, especially in embedded system (e.g., PDA and smart phone). In addition, it is known from a large amount of simulation that higher frequency resolution for noise-only frequency regions, which have usually low spectral magnitude, leads to the listener's fatigue.[1] Because of the above reasons, a method based on the warping of the input frequency without increasing the DCT size is proposed to adjust the spectral distribution of the input speech to be more appropriate for the DCT.

To warp the frequency axis, an all-pass transform that replaces $z^{-1}$ is proposed with a stable all-pass filter $A(z)$ defined by

$$A(z) = \frac{-\alpha + z^{-1}}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1 \tag{4}$$

where $\alpha$ is the control parameter for warping the frequency response, which is known as the *Laguerre* filter and is widely used in various signal processing algorithms [7], [9]. The resulting transfer function $F_k(A(z))$ is now an infinite impulse response (IIR) filter defined by

$$F_k\left(A(z)\right) = \sum_{n=0}^{M-1} U(k) \cos \frac{(2n+1)k\pi}{2M} \left(A(z)\right)^n. \tag{5}$$

### B. Implementation of WDCT

For the implementation, the filterbank method suggested in [7] is considered. When the filter is an $M$-tap finite impulse response (FIR) filter, the result of filtering and decimation by $M$ corresponds to the inner product of the filter coefficient vector and the input vector. From Parseval's relation, this is again equal to the inner product of the conjugate DFT of the input and the DFT of the filter coefficients which consists of the sampled values of $F_k(e^{j\omega})$ for $\omega = 0, (2\pi/M), (4\pi/M), \ldots, ((M-1)k\pi/M)$. Similarly, the result of filtering with $F_k(A(e^{j\omega}))$ is approximated by the inner product of the input vector and the inverse discrete Fourier transform (IDFT) of the sampled sequence of $F_k(A(e^{j\omega}))$. A more detailed description about the WDCT and its implementation can be found in [7].

The frequency responses of the warped filter banks for different values of $\alpha$ are shown in Fig. 1 where it can be seen that the low band is more emphasized with a positive $\alpha$. In contrast, a negative $\alpha$ is more appropriate for modeling the spectral characteristics in the high band. For that reason, in the case of the speech signal which mainly contains low-frequency components (i.e., voiced sounds), it is desirable to apply a positive value for $\alpha$. Similarly, a negative $\alpha$ is recommended for the speech signal which mostly has high-frequency components.

[1]In [14], lower frequency resolution with the merging of frequency bands is assigned to the high-frequency regions in which noise spectrum mainly locates for voiced sounds. It is known that the scheme is effective in reducing the disturbing noise in high-frequency parts for voiced sounds. However, this scheme does not work well for the speech signal with mostly high-frequency components.

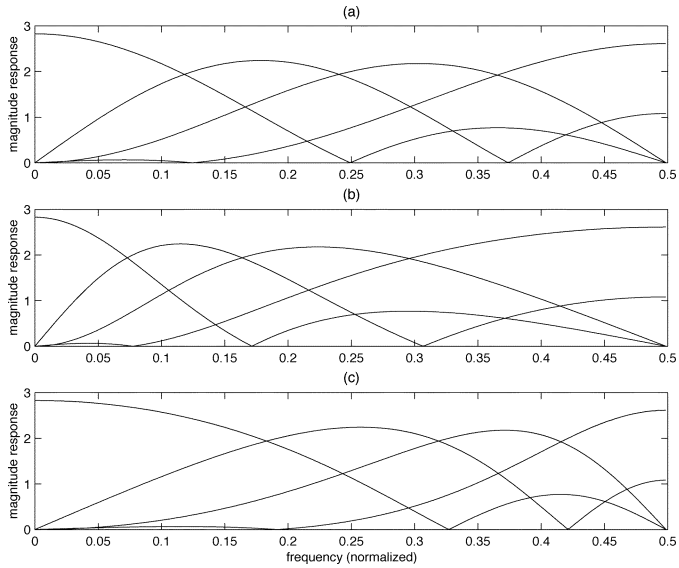


Fig. 1. Frequency responses for the four-point DCT/WDCT. (a) DCT filter bank. (b) WDCT filter bank with $\alpha = 0.25$. (c) WDCT filter bank with $\alpha = -0.25$.

## III. Frequency-Warped Speech Enhancement

It is assumed that a noise signal $n$ is added to a speech signal $x$, with their sum being denoted by $y$. Taking the $M$-point DCT gives us

$$Y_k(t) = X_k(t) + N_k(t), \qquad k = 0, 1, \cdots, M-1 \tag{6}$$

where $k$ denotes the $k$th frequency bin, $M$ is the total number of frequency components, and $t$ is the frame index in the time domain, respectively. Given a frame of noisy speech signal, the basic assumption adopted in our speech enhancement approach could be described by the following hypotheses:

$$H_0 : \text{speech absent} : \mathbf{Y}(t) = \mathbf{N}(t) \tag{7}$$

$$H_1 : \text{speech present} : \mathbf{Y}(t) = \mathbf{N}(t) + \mathbf{X}(t) \tag{8}$$

in which $\mathbf{Y}(t) = [Y_0(t), Y_1(t), \ldots, Y_{M-1}(t)]^T$, $\mathbf{N}(t) = [N_0(t), N_1(t), \ldots, N_{M-1}(t)]^T$, and $\mathbf{X}(t) = [X_0(t), X_1(t), \ldots, X_{M-1}(t)]^T$ are the DCT coefficients of the noisy speech, noise, and clean speech, respectively. The purpose of a speech-enhancement technique is to estimate $\{X_k(t), k = 0, 1, \ldots, M-1\}$ given $\{Y_k(t), k = 0, 1, \ldots, M-1\}$. Based on the Gaussian assumption, the MMSE estimator for $X_k$ is given by

$$\hat{X}_k = \frac{\xi_k}{1 + \xi_k} Y_k \tag{9}$$

where

$$\xi_k = \frac{\lambda_{s,k}}{\lambda_{n,k}} \tag{10}$$

in which $\lambda_{s,k}$ and $\lambda_{n,k}$ denote the variances of the clean speech and noise, respectively. The robust estimation of $\lambda_{n,k}$, $\lambda_{s,k}$, and $\xi_k$ also plays an essential role in the performance of speech enhancement. In this paper, the parameter estimation procedure proposed in [1] is adopted.

Specifically, multiplication in a transform domain corresponds to a filtering in time domain when the DCT is employed. Similarly, the linear convolution can be carried in the case of the WDCT [11]. It is noted that the MMSE estimator reduces to the Wiener filter when the real Gaussian assumption is adopted [10]. Additionally, spectral-domain-based speech enhancement such as the Wiener filter has a major drawback which is well known as "musical tone." Since this is a random frequency tone due to an underestimation of noise power, similar properties are made in a frequency-warped domain. To overcome this artifact, the soft-decision-based speech-enhancement algorithm is induced [1], [2]. For the extra configuration in speech enhancement, the basic framework proposed in [1] is adopted.

### A. Split-Band Global Soft Decision

In order to determine the frequency-warping control parameter $\alpha$, a statistical model is assumed for each split frequency band. For this, we first split the whole frequency range into high-band and low-band regions. Among the $M$ DCT coefficients, the leading $m$ coefficients are assigned to the low-band region while the remaining $(M-m)$ coefficients are used to form the high-band region. In the high-band region, the probability density functions (pdfs) of the noisy speech conditioned on $H_0$ and $H_1$ are assumed to be

$$p(Y_k|H_0) = \frac{1}{\sqrt{2\pi\lambda_{n,k}}} \exp\left\{-\frac{Y_k^2}{2\lambda_{n,k}}\right\} \tag{11}$$

$$p(Y_k|H_1) = \frac{1}{\sqrt{2\pi(\lambda_{s,k}+\lambda_{n,k})}} \exp\left\{-\frac{Y_k^2}{2(\lambda_{s,k}+\lambda_{n,k})}\right\}. \tag{12}$$

With the statistical assumptions shown in (11) and (12), the likelihood ratio $\Lambda(Y_k(t))$ is written as follows [5]:

$$\begin{aligned}\Lambda\left(Y_k(t)\right) &= \frac{p\left(Y_k(t)|H_1\right)}{p\left(Y_k(t)|H_0\right)} \\ &= \sqrt{\frac{\lambda_{n,k}}{\lambda_{s,k}+\lambda_{n,k}}} \exp\left\{-\frac{Y_k(t)^2}{2(\lambda_{s,k}+\lambda_{n,k})} + \frac{Y_k(t)^2}{2\lambda_{n,k}}\right\}.\end{aligned} \tag{13}$$

Applying the Bayes rule, we can easily derive the high-band global speech presence probability (HB-GSPP) such that

$$\begin{aligned}&P_H\left(H_1|\mathbf{Y}_H(t)\right) \\ &\quad= \frac{p(\mathbf{Y}_H(t)|H_1)\,P_H(H_1)}{p(\mathbf{Y}_H(t))} \\ &\quad= \frac{p(\mathbf{Y}_H(t)|H_1)P_H(H_1)}{p(\mathbf{Y}_H(t)|H_0)P_H(H_0)+p(\mathbf{Y}_H(t)|H_1)P_H(H_1)}\end{aligned} \tag{14}$$

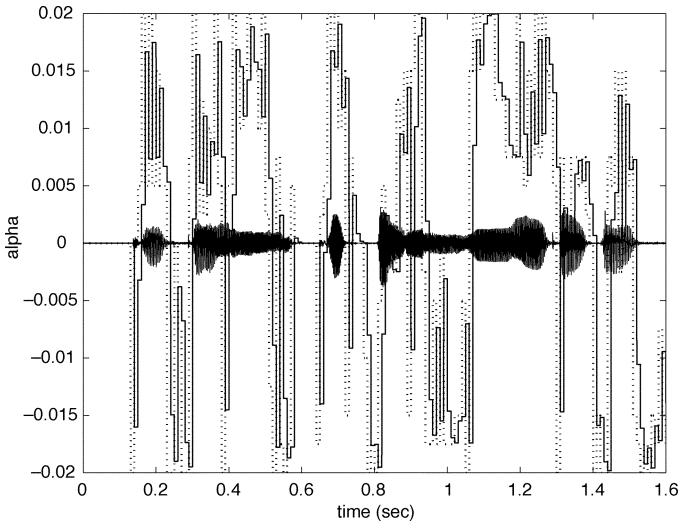


Fig. 2. Typical example of a trajectory of the $\alpha$ with corresponding speech waveform. A solid line for the $\hat{\alpha}(t)$ and a dotted line for the $\alpha(t)$ with $\lambda_p = 0.2$, respectively.

where $\mathbf{Y}_H(t) = [Y_m(t), Y_{m+1}(t), \ldots, Y_{M-1}(t)]$. Since the spectral component in each frequency bin is assumed to be statistically independent, (14) can be converted to

$$P_H\left(H_1|\mathbf{Y}_H(t)\right) = \frac{q_H \prod_{k=m}^{M-1} \Lambda\left(Y_k(t)\right)}{1 + q_H \prod_{k=m}^{M-1} \Lambda\left(Y_k(t)\right)} \tag{15}$$

where $q_H = P_H(H_1)/P_H(H_0)$.

The low-band global speech presence probability (LB-GSPP) is also computed in the same way to the computation of HB-GSPP such that

$$P_L\left(H_1|\mathbf{Y}_L(t)\right) = \frac{q_L \prod_{k=0}^{m-1} \Lambda\left(Y_k(t)\right)}{1 + q_L \prod_{k=0}^{m-1} \Lambda\left(Y_k(t)\right)} \tag{16}$$

where $\mathbf{Y}_L(t) = [Y_0(t), Y_1(t), \cdots, Y_{m-1}(t)]$ and $q_L = P_L(H_1)/P_L(H_0)$.

### B. Frequency-Warping Control Parameter Determination

In [7], the optimal frequency-warping control parameter is chosen so as to minimize the reconstruction error for the given image-compression algorithm. Also, to improve the speech recognition accuracy by compensating for the various vocal tract lengths, the optimal warping parameter should be determined for each speaker in a test or training set prior to the recognition stage in the speaker normalization [12], [13]. Since, however, these architectures cannot be applied directly to speech enhancement, it is necessary to choose the efficient warping parameter in an online fashion. So, the determination of the frequency-warping control parameter $\alpha$ based only on the input speech in each frame should be considered. A straightforward way one can consider this is to apply the voiced/unvoiced (V/UV) decision and select $\alpha$ depending on the decision, i.e., a positive value for the voiced sound and a negative value for the unvoiced speech with high energy in the high-band region.

In this section, a method to determine $\alpha$ by using only the HB-GSPP and LB-GSPP is proposed. This method is based on the assumption that the positive $\alpha$ is chosen for the voiced sound

TABLE I
MOS RESULTS FOR THE PROPOSED WDCT AND DCT-BASED SPEECH-ENHANCEMENT METHODS

| noise | babble | | | car | | | white | | | street | office |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SNR(dB) | 5 | 10 | 15 | 5 | 10 | 15 | 5 | 10 | 15 | approx. 15 | approx. 15 |
| none | 1.89 | 2.52 | 2.75 | 2.56 | 3.29 | 3.66 | 1.05 | 1.51 | 2.00 | 3.25 | 2.37 |
| DCT | 2.43 | 3.11 | 3.50 | 2.70 | 3.58 | 3.90 | 1.84 | 2.49 | 3.11 | 3.51 | 2.54 |
| WDCT | 2.69 | 3.39 | 3.72 | 3.14 | 3.78 | 4.05 | 2.40 | 2.95 | 3.44 | 3.73 | 2.64 |

TABLE II
SEGMENTAL SNR RESULTS FOR THE PROPOSED WDCT- AND DCT-BASED SPEECH-ENHANCEMENT METHODS

| noise | babble | | | car | | | white | | |
|---|---|---|---|---|---|---|---|---|---|
| segmental SNR(dB) | 5 | 10 | 15 | 5 | 10 | 15 | 5 | 10 | 15 |
| DCT | 8.68 | 12.62 | 16.98 | 9.28 | 13.45 | 17.93 | 10.81 | 14.38 | 18.20 |
| WDCT | 8.86 | 12.83 | 17.09 | 9.30 | 13.47 | 17.93 | 11.12 | 14.74 | 18.53 |

which is more concentrated in the low-band region while the negative $\alpha$ is selected for the speech signal which has most of its energy in the high-band region. Based on the soft-decision scheme which is known to be more helpful to avoid an abrupt discontinuity in spectral components [6], the proposed method is described in the following way:

$$\alpha(t)=\begin{cases}\alpha_{\min}\cdot P_H(H_1|\mathbf{Y}_H(t)), & \text{for } P_H(H_1|\mathbf{Y}_H(t))>P_{\min} \\ & P_L(H_1|\mathbf{Y}_L(t))<P_{\min} \\ \alpha_{\max}\cdot P_L(H_1|\mathbf{Y}_L(t)), & \text{for } P_L(H_1|\mathbf{Y}_L(t))>P_{\min} \\ & P_H(H_1|\mathbf{Y}_H(t))<P_{\min} \\ 0, & \text{elsewhere}\end{cases} \tag{17}$$

where $P_{\min}=0.2$ and $\alpha\in[\alpha_{\min}(=-0.02),\alpha_{\max}=(0.02)]$. The values $\alpha_{\max}$ and $\alpha_{\min}$ are determined based on a variety of experimental tests. According to the experiment, since higher values of $\alpha$ lead to a signal degradation, an experimentally optimized value is chosen.[2] Considering (17), it is not difficult to find out that $\alpha(t)$ becomes $\alpha_{\min}$ as HB-GSPP approaches one only if LB-GSPP is sufficiently small. On the other hand, $\alpha(t)$ approaches $\alpha_{\max}$ as LB-GSPP increases while HB-GSPP is kept low. A typical example of trajectory of $\alpha$ is reported by Fig. 2. From the result, it is evident that $\alpha$ is positive for the voiced periods while it is negative during the speech parts with mostly high-frequency components.

For the purpose of avoiding a rapid variation, a temporal smoothing technique to $\alpha(t)$ is applied such that

$$\hat{\alpha}(t+1)=\lambda_p\cdot\hat{\alpha}(t)+(1-\lambda_p)\cdot\alpha(t) \tag{18}$$

where $\hat{\alpha}(t)$ denotes the smoothed control parameter and $\lambda_p$ is a smoothing parameter. In order to implement the WDCT-based speech-enhancement technique, a WDCT matrix for each value of $\alpha$ is necessary. Since the computation of the WDCT matrix for a specific value of $\alpha$ requires a large computation, it is beneficial to precompute the WDCT matrices and store them. For this, $[\alpha_{\min},\ \alpha_{\max}]$ is divided uniformly into 16 regions and a WDCT matrix is constructed for each region with the center value. For reducing the memory size while taking into account the time-invariant characteristics of the speech signal, it is better to quantize $\alpha$ into 16 steps.

[2]These values are smaller than $[-0.1,\ 0.1]$ of Cho *et al.* in [7] since it is observed that smaller values are preferred in speech enhancement.

During speech enhancement, the region to which $\hat{\alpha}(t)$ belongs is identified, and the WDCT matrix corresponding to that region is applied to transform the data.

## IV. EXPERIMENTS AND RESULTS

This section presents the performance of the proposed WDCT-based speech-enhancement approach as well as a comparison with the DCT-based one. For verifying the performance of the proposed approach, not only the objective quality measurement but also the subjective quality evaluation test was carried out. Let us explain the experimental environments for the comparison test. Eight test sentences, in which four were spoken by a male speaker and the others were generated by a female speaker, were sampled at 8 kHz and used for evaluation. A trapezoidal window of length 13 ms was applied to the input signal every 10 ms which is similar to the noise suppression rule in the IS-127 standard. By overlapping adjacent frames (3 ms), the blocking effect (block discontinuity) of the DCT can be reduced [10]. Each frame of the windowed signal was transformed to the corresponding spectrum through 128-point WDCT after zero padding frame by frame.[3]

For the split-band approach, we use $m=(2/3)M$ which is an experimentally chosen value for the robust determination of *HB-GSPP* and *LB-GSPP* in the aforementioned soft-decision-based scheme. Three types of noise sources—the babble, white, and car noises from the NOISEX-92 database—were electrically added to the clean speech waveforms by varying the SNR. At first, several MOS tests on a number of enhanced noisy speech samples based on the noise environments were conducted. Furthermore, we consider acoustical noisy speech, where the speech signal is recorded in real noise conditions which are in the street and office for the MOS tests. The listening tests were performed by ten listeners and each listener gave a score from one to five for each test sequence. This score represents his or her global appreciation of the residual noise and the speech distortion. The MOS results are shown in Table I where, for the purpose of comparison, we also list the result provided by the enhancement technique based on the conventional DCT. According to [1], our previous DCT-based enhancement technique showed clear improvements compared with the IS-127 noise suppression rule [14]. Actually, the IS-127 noise suppression employs the mel-scaled filter bank.

[3]For $t=0$, $\alpha=0$.

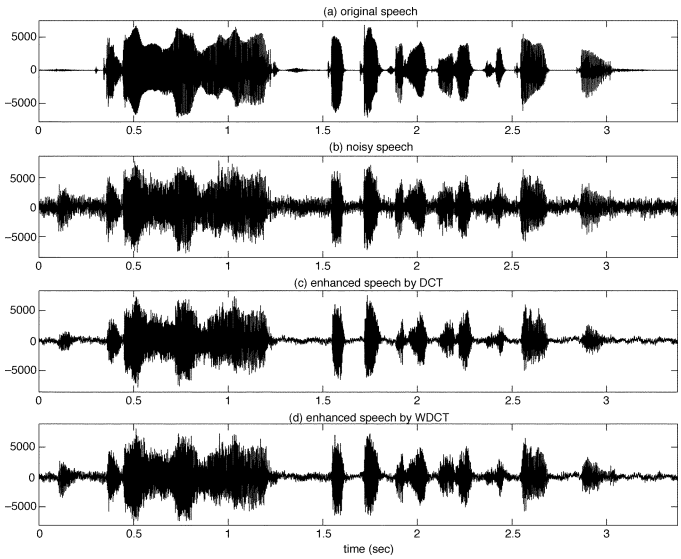


Fig. 3. Comparison of speech segment under the babble noise at SNR $= 5$ dB. (a) Clean speech. (b) Noisy speech. (c) Enhanced speech by DCT. (d) Enhanced speech by WDCT.

This differs from our proposed WDCT-based enhancement scheme which warps frequency axis into higher or lower frequency regions while the IS-127 uses only the combination of the frequency channel based on the mel-scale. From the MOS results, it can be seen that, in most noise conditions, the proposed WDCT-based method yielded higher scores than the DCT-based algorithm.

Second, in order to evaluate the objective quality, a segmental SNR was considered as it is defined by

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{\sum_{t=0}^{N-1} \mathrm{SNR}(t)}{N} \tag{19}$$

where $\mathrm{SNR}(t)$ represents the SNR computed in the $t$th frame, and $N$ is the total number of frames in the given data. The results for the segmental SNR shown in Table II. In all test conditions, the proposed WDCT-based algorithm significantly outperformed or at least was comparable to the DCT-based one. It is noted that the improvement in MOS scores is more significant than that in segmental SNR since the parameters have been optimized for subjective quality. Fig. 3 illustrates the clean, noisy speech under the babble $(\mathrm{SNR} = 5\ \mathrm{dB})$ and the results of enhancement using the different algorithm (DCT or WDCT) for easy understanding of the performance difference.

## V. Conclusion

In this paper, the WDCT-based speech-enhancement technique is proposed. WDCT is considered as a cascade of an adjustable all-pass IIR filter and the conventional DCT, which results in an adaptive transform of the input speech. The warping control parameter is determined based on split-band analysis. The performance of WDCT has been found to be much better than that of the conventional DCT with a small additional computation burden since WDCT matrices are predefined in a prescribed set of frequency ranges.

## Acknowledgment

The author would like to thank the anonymous reviewers for helpful suggestions and comments.

## References

[1] J.-H. Chang and N. S. Kim, "Speech enhancement: new approaches to soft decision," *IEICE Trans. Inf. Syst.*, vol. 27, pp. 1231–1240, Sep. 2001.

[2] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Process. Lett.*, vol. 7, no. 5, pp. 108–110, May 2000.

[3] R. J. McAulary and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 137–145, Apr. 1980.

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

[5] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Commun.*, vol. 24, no. 3, pp. 249–257, 1998.

[6] J.-H. Chang and N. S. Kim, "Speech enhancement using warped discrete cosine transform," in *Proc. IEEE Speech Coding Workshop*, Tsukuba, Japan, Oct. 2002.

[7] N. I. Cho and S. K. Mitra, "Warped discrete cosine transform and its application in image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 12, pp. 1364–1373, Dec. 2000.

[8] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.

[9] A. Markur and S. K. Mitra, "Warped discrete-Fourier transform: Theory and applications," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 48, no. 9, pp. 1086–1093, Sep. 2001.

[10] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithm, Advantages, Applications*. New York: Academic, 1990.

[11] S. Bagchi and S. K. Mitra, *The Nonuniform Discrete Fourier Transfrom and Its Applications in Signal Processing*. Norwell, MA: Kluwer, 1999.

[12] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedure," in *Proc. IEEE Int. Conf. Acoustics. Speech, and Signal Processing*, vol. 1, Atlanta, GA, May 1996, pp. 339–341.

[13] J. McDonough, W. Byrne, and X. Luo, "Speaker normalization with all-pass transforms," in *Proc. Int. Conf. Spoken Language Processing*, vol. 6, Sydney, Australia, Nov. 1998, pp. 2307–2310.

[14] *Enhanced Variable Rate Codec, Speech Service Option 3 for Wide-band Spectrum Digital Systems*, 1996.