WARPED DISCRETE COSINE TRANSFORM CEPSTRUM: A NEW FEATURE FOR SPEECH PROCESSING

R. Muralishankar[1], *Abhijeet Sangwan*[2] *and Douglas O'Shaughnessy*[1]

[1]INRS-EMT (Telecommunications), University of Quebec, Montreal, Canada.
[2]Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada.
murali@inrs-emt.uquebec.ca, a_sangwa@ece.concordia.ca, dougo@inrs-emt.uquebec.ca

ABSTRACT

In this paper, we propose a new feature for speech recognition and speaker identification application. The new feature is termed as *warped-discrete cosine transform cepstrum* (WDCTC). The feature is obtained by replacing the discrete cosine transform (DCT) by the warped discrete cosine transform (WDCT, [4]) in the discrete cosine tranform cepstrum (DCTC [2]). The WDCT is implemented as a cascade of the DCT and IIR all-pass filters. We incorporate a nonlinear frequency-scale in DCTC which closely follows the bark-scale. This is accomplished by setting the all-pass filter parameter using an expression given by Smith and Abel [5] . Performance of WDCTC is compared to mel-frequency cepstral coefficients (MFCC) in a speech recognition and speaker identification experiment. WDCTC outperforms MFCC in both noisy and noiseless conditions.

1. INTRODUCTION

For the past two decades, the feature extraction algorithm of Davis and Mermelstein [1] has been used extensively in the field of speech and audio processing. The algorithm stems from two ideas: (1) vocal tract modeling, and (2) homomorphic filtering. In the vocal tract model, speech is produced by passing an excitation through a filter whose response models the effects of the vocal tract on the excitation. In homomorphic filtering, the convolution of the excitation with the vocal tract response is transformed to addition, where linear filtering techniques are applied to remove the excitation from the filter response. Unlike homomorphic filtering, the Davis and Mermelstein algorithm generates mel-frequency cepstral coefficients (MFCC) by transforming the log-energies of the spectrum passed through a bank of band-pass filters. The MFCC filter bank is composed of triangular filters spaced on a logarithmic scale. The spacing of the filters follow the mel-scale, which is inspired by the critical band measurements of the human auditory system.

MFCC was compared to the linear predictive cepstral coefficients (LPCC) and the discrete cosine transformed cepstrum (DCTC [2]) in a speaker identification task presented by Muralishankar et al [3]. Speaker identification using DCTC was shown to perform better than LPCC but worse than MFCC [3]. It is observed that the underperformance of DCTC when compared to MFCC can be attributed to the absence of a nonlinear frequency scale. We propose to incorporate the nonlinear frequency scale for the DCTC feature which we believe will improve its performance to match that of MFCC. This is achieved by using the warped discrete cosine transform (WDCT) instead of DCT in obtaining DCTC. The WDCT is implemented as a cascade of the DCT and IIR all-pass filters whose parameters are used to adjust the transform according to the frequency contents of the signal block [4]. As a parallel, the advantage of using WDCT over DCT for image compression has been shown in [4].

Smith and Abel [5] derived an analytic expression for an all-pass filter parameter such that the mapping between the warped and the unwarped frequencies, for a given sampling frequency f_s, follows the psychoacoustic Bark-scale. This expression is used to get WDCT. Applying WDCT to DCTC generates a new feature, warped discrete cosine transform cepstrum (WDCTC). To evaluate the efficacy of the new feature in speech processing applications, we compare the performances of WDCTC with MFCC for vowel recognition and speaker identification tasks. Our experimental results show that WDCTC consistently performs better than MFCC.

This paper is organized as follows. In section 2, we introduce the WDCT and present its nonlinear frequncy resolution property. In section 3, WDCT is incorporated into DCTC. In section 4, we present the comparative performances of WDCTC and MFCC for vowel recognition and speaker identification tasks. Section 5 provides the concluding remarks.

2. WARPED DISCRETE COSINE TRANSFORM

2.1 Definition

Here, we review an N-point WDCT of the input vector $[x(0),x(1),...,x(N-1)]^T$[4]. The N-point DCT, $\{X(0),X(1),...,X(N-1)\}$ is defined by

$$X(k) = U(k)\sum_{n=0}^{N-1} x(n)\cos\frac{(2n+1)k}{2N}\pi \tag{1}$$

for $k = 0,1,...,N-1$ where

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

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

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

That is, the i^{th} coefficient of $F_k(z^{-1})$ is the $(k,i)^{th}$ element of the DCT matrix. It can be shown that $F_k(z^{-1})$ is a band-pass filter with a center frequency at $(2k+1)/2N$, with the sampling frequency normalized to 1. Hence, the magnitude response of $F_k(z^{-1})$ for small k is larger for low-frequency inputs such as voiced sounds, which enable data compression by giving more emphasis to the lower band outputs than the higher band ones [4]. Further, inputs with mostly high-frequency components such as unvoiced sounds have a higher magnitude response of $F_k(z^{-1})$ for large k, which enables high frequency coefficients to have compacted energy. This is a desirable feature for noise removal purposes [6].

Note that the frequency resolution of the DCT is uniform. Therefore, incorporating a nonlinear frequency resolution closely following the psychoacoustic Bark-scale will result in an enhanced representation for the speech signals. We introduce such a nonlinearity in DCTC using warping. To warp the frequency axis, we apply an all-pass transformation by replacing z^{-1} with an all-pass filter $A(z)$ defined by

$$A(z) = \frac{-\beta + z^{-1}}{1 - \beta z^{-1}} \qquad (4)$$

where β is the control parameter for warping the frequency response. $A(z)$ is known as the Laguerre filter and is widely used in various signal processing algorithms. The resulting $F_k(A(z))$ now becomes an infinite impulse response (IIR) filter given by

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

2.2 Implementation of the WDCT

The WDCT can be implemented in several ways. The most straightforward approach is to implement the filters in a Laguerre network (considering first order all-pass filters, $A(z)$, which are reset every N samples). In the second approach, we can implement the filtering by a matrix-vector multiplication in two steps: first we divide the all-pass IIR transfer functions into N terms, and then sample the frequency responses of the warped filter bank to obtain the WDCT matrix through an inverse discrete Fourier transform (IDFT).

We use the second approach, which is the filter bank method suggested by Cho and Mitra [4]. For an N-tap finite impulse response (FIR) filter, the result of filtering and decimation by N 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 is equal to the sampled value of $F_k(e^{j\omega})$ for $\omega = (2\pi k/N)$ where $k = 0, 1, ..., N-1$. Similarly, we can approximate the result of the filtering with $F_k(A(e^{j\omega}))$ as the inner product of the input vector and the IDFT of the sampled sequence of $F_k(A(e^{j\omega}))$. More detailed description about the WDCT and its implementations can be found in [4].

2.3 Selection of the all-pass filter coefficient

Nonuniform resolution fast Fourier transform (FFT) was introduced by Oppenheim, et al [7]. The main idea was to use a network of cascaded first order all-pass filter sections for frequency warping of the signal and then apply FFT to produce the warped spectrum from the preprocessed signal.

The transfer function of a first order all-pass filter is given in eq. 4. By definition, the magnitude response of the filter is a constant. The phase response of $A(z)$ is given by

$$\varpi = \omega + 2\arctan\left(\frac{\beta\sin(\omega)}{1 - \beta\cos(\omega)}\right). \qquad (6)$$

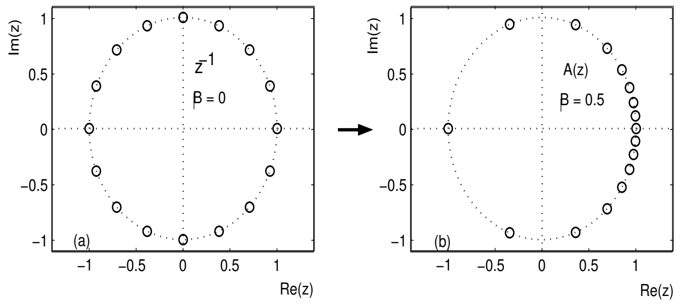

Figure 1: Linear and warped frequency sampling. (a) Linear frequency sampling of trivial all-pass filter ($\beta = 0$). (b)Warped frequency resolution of first-order all-pass filter ($\beta = 0.5$).

The phase function determines a frequency mapping occuring in the all-pass chain [7]. Depending on the sign of β, the low or high frequency range is expanded whereas the remaining part of the unit circle becomes compressed. This is shown in Fig. 1. For a certain value of β the frequency transformation closely resembles the frequency mapping occuring in the human auditory system. Smith and Abel [5] derived an analytic expression for β so that the mapping, for a given sampling frequency f_s, matches the psychoacoustic Bark-scale mapping. The value is given by

$$\beta \approx 1.0211\left(\frac{2}{\pi}\arctan(0.076 f_s)\right)^{\frac{1}{2}} - 0.19877. \qquad (7)$$

For a given f_s (say 16 kHz), we calculated β from eq. 7 to generate the WDCT matrix.

3. WARPED DISCRETE COSINE TRANSFORMED CEPSTRUM

Two variants of DCTC were proposed in [3], namely, DCTC-1 and DCTC-2. It was shown that DCTC-2 performs better than DCTC-1 and both outperform LPCC in a speaker identification task. Hence, we chose the DCTC-2 algorithm and replaced DCT with WDCT to obtain the WDCTC algorithm. The new WDCTC algorithm is outlined below.

Consider a finite duration, real sequence $x(n)$, defined for $0 \le n \le N-1$ and zero elsewhere. Taking an N-point WDCT of the above sequence, we have $X_{WDCT}(k)$ defined for $0 \le k \le N-1$. We can write $X_{WDCT}(k)$ as

$$X_{WDCT}(k) = \exp(\xi(k))\,|X_{WDCT}(k)| \qquad (8)$$

where

$$\xi(k) = \frac{j\pi}{2}(\mathrm{sgn}(X_{WDCT}(k)) - 1)$$

and

$$\mathrm{sgn}(p) = \begin{cases} 1, & \text{for} \quad p \ge 0 \\ -1, & \text{for} \quad p < 0 \end{cases}.$$

Taking natural logarithm on both sides of eq. 8,

$$\ln\{X_{WDCT}(k)\} = \xi(k) + \ln|X_{WDCT}(k)|, \qquad (9)$$

then we obtain the WDCTC of $x(n)$ as

$$\widehat{x}(n) = \mathrm{Re}\{IDCT(\xi(k) + \ln|X_{WDCT}(k)|)\}. \qquad (10)$$

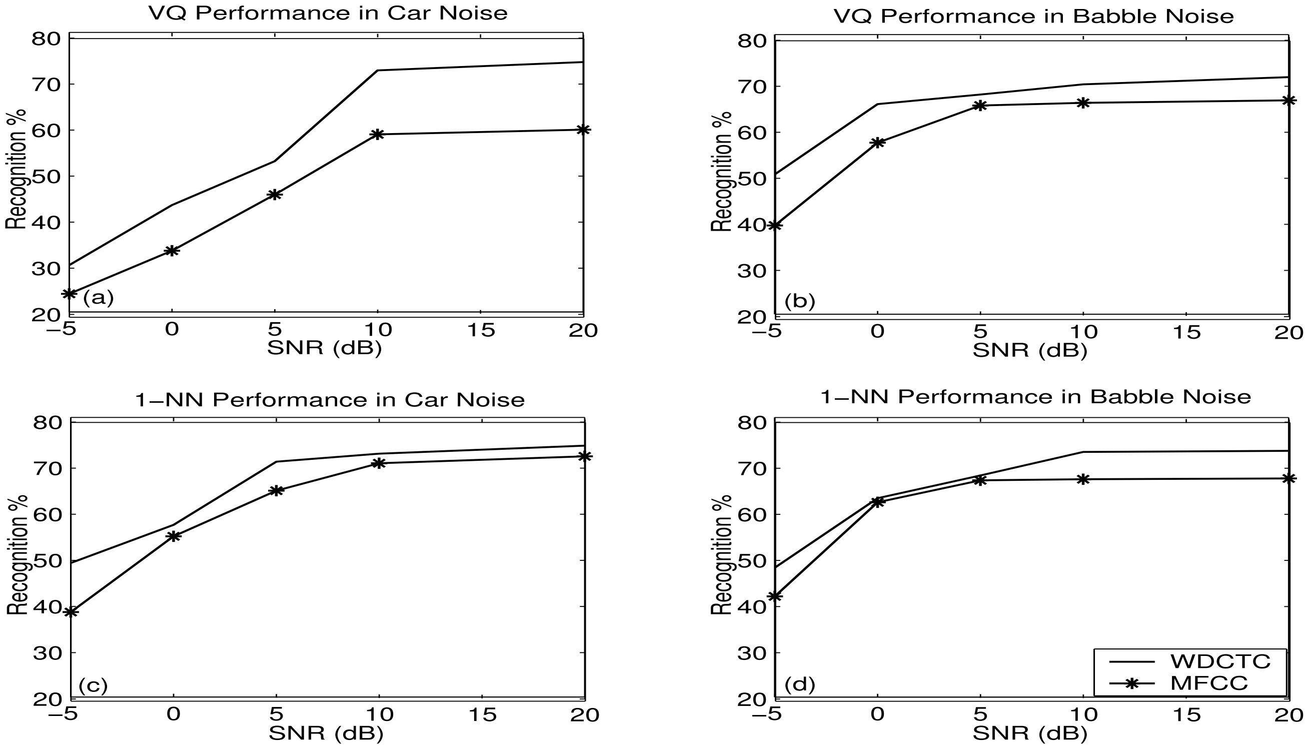

Figure 2: Noisy vowels average recognition performances of MFCC and WDCTC. (a) Using VQ model in presence of CAR noise. (b) Using VQ model in presence of Babble noise. (c) Using 1-NN model in presence of CAR noise. (d) Using 1-NN model in presence of Babble noise.

Here, an inverse discrete cosine transform (IDCT)[8] is used to get the WDCTC sequence and it is denoted as $\widehat{x}(n)$.

4. RESULTS AND DISCUSSION

To test the effectiveness of our proposed feature WDCTC, we conducted two experiments: 1. vowel recognition, 2. speaker identification. We compare the recognition performance of the WDCTC feature with MFCC.

4.1 Vowel recognition

Vowel recognition experiments were conducted on the TIMIT database. We selected 5 vowels $/aa/$,$/eh/$,$/iy/$,$/ow/$ and $/uw/$ for our experiments. The vowel segments were extracted from continuous speech using the train and test set of the TIMIT database (Dialect Region: North Midland) to form the train and test set for our experiments. The number of speakers in our train and tests sets were 72 and 26, respectively.

Each vowel segment was sampled at 16 kHz. Duration of each frame of speech was 16 ms, with an overlap of 8 ms between successive frames. Each frame of speech was preemphasized with a factor equal to 0.98 and Hamming windowed. Eighteen dimensional feature vectors (MFCC and WDCTC) were obtained for each frame. For obtaining MFCC, the Mel-scale was simulated using a set of 18 triangular filters. For WDCTC, we used eq.7 to get corresponding β for a given f_s. The WDCT matrix was generated using this β. The WDCTC feature was generated by using eq. 10. The first 18 WDCTC coefficients, excluding the gain term, form the feature vector.

Each vowel is modeled using a 32-length vector quantization (VQ) codebook, consisting of code vectors of MFCC and WDCTC coefficients. The codebooks are trained using the k-means clustering algorithm [11], employing a Euclidean distance measure. Vowels are identified by evaluating the distortion between the features of the test vowel sample and the models in the vowel database. We have used two classifiers for our experiments: VQ and 1-Nearest Neighbour (NN) model [13].

Car noise and Babble noise were added to noiseless vowels. To simulate different noisy conditions, the variance of the noise was adjusted with respect to the variance of noiseless vowels to obtain target signal-to-noise ratios (SNR) for noisy vowels. SNR was varied from -5 dB to 20 dB and the noisy vowel recognition performances using MFCC and WDCTC features were obtained. The vowel recognition performance of MFCC and WDCTC features for clean speech using VQ and the 1-NN model is shown in Table 1. Figure 2 shows the noisy vowel recognition performances of MFCC and WDCTC. Figure 2(a) and (b) show the recognition accuracy in percentage and the comparative performances in the presence of car noise and babble noise for the VQ model. Similarly, Figs. 2(c) and (d) show the results for 1-NN model. From Table 1 and Fig. 2, we can see that WDCTC outperforms MFCC.

4.2 Speaker Identification

We employed WDCTC as a feature for text-independent speaker identification. Our experiments with WDCTC and MFCC demonstrate the viability of WDCTC for speaker identification. We have performed speaker identification ex-

Table 1: Comparision of clean vowel average recognition performance of MFCC and WDCTC features.

	VQ model		1-NN model	
Feature	Train	Test	Train	Test
MFCC	90.86	67.83	99.01	71.51
WDCTC	90.97	69.48	99.75	73.00

periments on a limited subset (21 male speakers) of the NIST evaluation corpus 1996 database [9]. This corpus was derived from the entire switchboard-I corpus for speaker verification evaluations. The development data consisted of training and test speech from 45 male and 45 female speakers which was sampled at 8 kHz. The evaluation data consisted of training data from 21 male and 19 female target speakers plus test speech from these targets and 167 male and 216 female unseen imposters [10]. There are 3 training conditions for each target speaker. These conditions are 1). one-session training 2). one-handset training and 3). two-handset training. All these conditions use 2 minutes of training speech data from the target speaker.

We intend to compare the speaker-identification performances of WDCTC and MFCC rather than use these features for speaker verification. So, we chose one-session and one-handset conditions for training. Further for testing, we used one (Test Set 1) and two (Test Set 2) handset conditions from the training set of NIST. We obtain the 18 dimensional MFCC and WDCTC feature vectors for speaker train and test sets after excluding the first coefficients from both. Each speaker is modeled using a 32-length vector quantization codebook, consisting of code vectors of MFCC and WDCTC coefficients. The codebooks are trained using the k-means clustering algorithm [11], employing a euclidean distance measure. Speakers are identified by evaluating the distortion between the features of the test speech sample and the models in the speaker database. We use VQ and 1-NN [13] for classification. The experiments are evaluated as in [12]. The test speech of each speaker is processed to produce a sequence of feature vectors. This sequence is then divided into overlapping segments of 100 feature vectors each, with an overlap of 90 feature vectors between successive segments. Each segment is considered a separate test utterance. If N_T is the total number of segments and N_C is the number of correctly identified segments, then the identification performance is evaluated in percentage as $\frac{N_C}{N_T} \times 100$. Thus, a segment based performance metric, rather than direct speaker recognition performance, has been used. The speaker identification results are shown in Table 2. We can see better performance of WDCTC over MFCC. Certainly, the overall accuracy achieved for this task is low. This is due to the fact that we are using simple classifiers and mainly bank on the feature itself for the performance. Further, the recognition accuracy of Test Set 1 is high as the test set has been taken from the training data itself.

5. CONCLUSION

We have presented WDCTC as a feature for vowel recognition and speaker identification. In our experiments, the performance using WDCTC is consistently better than MFCC. The improved performance of WDCTC may be due to incorporation of nonlinear frequency-scale which approximates the bark-scale. The added binary phase in WDCTC may also contribute towards better performance over MFCC. Here, the classifiers (VQ model and 1-NN) are chosen to minimize their influence over the features. However, the results thus obtained are low compared to the state of the art in vowel recognition and speaker identification tasks.

Table 2: Comparision of speaker identification performance of MFCC and WDCTC features.

	Test Set 1		Test Set 2	
Speaker Identification Rate (%) using	MFCC	WDCTC	MFCC	WDCTC
1-NN Model	90.11	96.32	23.5	29.63
VQ Model	27.05	33.57	2.75	8.99

REFERENCES

[1] Steven B. Davis and Paul Mermelstein, “Comparision of parametric representation for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 28(4), pp. 357-366, 1980.

[2] R. Muralishankar and A. G. Ramakrishnan, “DCT based Pseudo-complex cepstrum,” *ICASSP’02,* vol. 1, pp. 521-524, 2002.

[3] R. Muralishankar and A. G. Ramakrishnan, “Pseudo complex cepstrum using discrete cosine transform,” accepted, *International Journal of Speech Technology*.

[4] 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, pp. 1364-1373, Dec. 2000.

[5] J. O. Smith III and J. S. Abel, “Bark and ERB Bilinear Transforms,” *IEEE Trans. Speech, Audio Processing,* vol. 7, pp. 697-708, June 1999.

[6] I. Y. Soon, S. N. Koh and C. K. Yeo, “Noisy speech enhancement using discrete cosine transform,” *Speech communication,* vol. 24, No. 3, pp. 249-257, 1998.

[7] A. V. Oppenheim, D. H. Jhonson, K. Steiglitz, “Computation of spectra with unequal resolution using the FFT,” *Proc. IEEE,* vol. 59, pp. 299-301, Feb. 1971.

[8] S. A. Matrucci, “Symmetric convolution and the discrete sine and cosine transforms,” *IEEE Transactions on Signal Processing*, vol. 42, pp. 1038-1051, 1994.

[9] NIST Speaker recognition evaluations plan, http://www.nist.gov/speech/test.htm

[10] J. P. Campbell, Jr. and D. A. Reynolds, “Corpora for the evaluation of speaker recognition systems,” *ICASSP-99*, vol. 2, pp. 829-832, 1999.

[11] R. Duda, P. Hart and D. G. Stark, *Pattern classification*, J. Wiley, New York, 2002.

[12] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech, Audio Processing,* vol. 3, pp. 72-83, 1995.

[13] Vishwa N. Gupta, Matthew Lennig, and Paul Mermelstein, “Decision rules for speakerindependent isolated word recognition,” *ICASSP ’84*, vol. 9, pp. 336-339, 1984.