基於離散餘弦轉換之語音特徵的強健性補償法
Compensating the speech features via discrete cosine transform for robust speech recognition

Hsin-Ju Hsieh 謝欣汝, Wen-hsiang Tu 杜文祥, Jeih-weih Hung 洪志偉
暨南國際大學電機工程學系
Department of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
E-mail: s98323550@ncnu.edu.tw, aero3016@ms45.hinet.net, jwhung@ncnu.edu.tw

Abstract

In this paper, we develop a series of algorithms to improve the noise robustness of speech features based on discrete cosine transform (DCT). The DCT-based modulation spectra of clean speech feature streams in the training set are employed to generate two sequences representing the reference magnitudes and magnitude weights, respectively. The two sequences are then used to update the magnitude spectrum of each feature stream in the training and testing sets. The resulting new feature streams have shown robustness against the noise distortion. The experiments conducted on the Aurora-2 digit string database reveal that the proposed DCT-based approaches can provide relative error reduction rates of over 25% as compared with the baseline system using MVN-processed MFCC features. Experimental results also show that these new algorithms are well additive to many noise robustness methods to produce even higher recognition accuracy rates.

I. Introduction

Most of the state-of-the-art automatic speech recognition (ASR) system developed in the laboratory, in which the speech is not obviously distorted, can achieve excellent recognition performance. But in the real-world application, the recognition accuracy is seriously degraded due to so many distortions or variations existing in the application environment. Particularly speaking, the environmental distortions can be roughly classified into two types: channel distortion and additive noise, both influencing the performance of an ASR system a lot. The channel distortion occurs when the speech signal is transmitted by electronic devices or transmission lines, such as the air, the telephone line or the microphone. The additive noise is like the “shadow” or “background” existing in the environment, such as car noise and babble noise. Noise robustness techniques have thus received much attention in recent years since they are so important in the applicability of ASR.

One school of noise-robustness techniques is devoted to compensate the original speech feature to reduce the effect of noise and recover the speech feature back to its intact state. Typical examples of these techniques include cepstral mean normalization (CMN) [1], mean and variance normalization (MVN) [2], cepstral gain normalization (CGN) [3], cepstral shape normalization (CSN) [4], histogram equalization (HEQ) [5], higher-order cepstral moment normalization (HOCMN) [6], temporal structure normalization (TSN) [7] and MVN plus ARMA filtering (MVA) [8]. However, the main purpose of the above methods can be roughly divided into two parts: one is to normalize the statistics of temporal-domain feature sequence and the other is to further reduce the mismatch by enhancing some components which are not easily affected by noise. For the latter case, the discrete Fourier transform (DFT) is usually used to be an analysis tool for obtaining the modulation spectrum of temporal-domain feature sequence. Therefore, we can deal with the modulation spectrum explicitly or implicitly in order to obtain the robust temporal-domain feature sequence.

In this paper, we present two novel methods to improve the noise robustness of speech features, hoping to promote the resulting recognition accuracy. These novel methods take advantage of the discrete cosine transform (DCT) [9] to analyze and cope with the temporal-domain feature sequence, which is quite different form the conventional DFT-based methods. As we know, DCT is widely used in many fields, such as image compressing and coding. However, it is less used for robust speech feature extraction. Especially, to our knowledge, there are little research that directly uses DCT to analyze and process the temporal-domain feature sequence. Therefore, the proposed methods in this paper are both innovative and valuable.

The remainder of the paper is organized as follows: Section II describes an overview of DCT and the effect of noise on the DCT-based modulation spectrum of speech features. Then the details of our proposed feature compensation algorithms based on DCT are described in Section III. Section IV contains the experimental setup, experimental results and discussions. Finally, concluding remarks are given in Section V.

II. Brief introduction of discrete cosine transform (DCT) and the effect of noise on the DCT of the speech feature streams

Discrete cosine transform (DCT) is a Fourier-related transform similar to discrete Fourier transform (DFT), and it has been one of the most powerful analysis tools in the field of signal processing. Basically speaking, DCT expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. DCT has been successfully applied in many aspects of speech analysis, like transform coding and speech feature extraction. It transforms the input signal from the time domain into the frequency domain, which highlights the periodicity of the signal. Besides, in speech feature extraction, DCT plays an important role in reducing the correlation of features and thus results in a more compact feature representation. In the following, we will make a brief introduction of DCT, and then investigate the effect of noise on the DCT of the speech feature stream, which serves as the background of the presented methods in section III.

II.1 The relationship between DCT and DFT

DCT expresses a signal in terms of a weighted sum of sinusoids, which is similar to DFT. However, DCT has some peculiar properties that are different from DFT. An obvious distinction between DFT and DCT is that, in analyzing a real-valued signal, DFT uses complex sinusoids (including the cosine and sine functions), while the latter uses only cosine functions. As a result, DFT often exhibits complex values while DCT real values only, indicating that the DCT coefficients are either 0 (positive) or π (negative) in phase.

It can be shown that the DCT of a signal $x[n]$ equals to the amplitude part of the DFT of another signal $y[n]$ given $y[n]$ is an extended version of $x[n]$ with even symmetry. According to different arrangements for the even-symmetry condition, eight DCT variants can be defined, among which the type-II DCT is probably the most commonly used form, and is often simply referred to as “the DCT”. Besides, the inverse of the type-II DCT (IDCT) is just the type-III DCT.

For a finite-length real-valued sequence $\{x[n];\ 0 \le n \le N-1\}$, its DFT $X[k]$ and DCT (type-II DCT) $C[k]$ are obtained by the following two equations, respectively:

$$\textbf{DFT:}\ \ X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi kn}{N}},\ \ 0 \le k \le N-1, \tag{1}$$

$$\textbf{DCT:}\ \ C[k] = \frac{1}{\sqrt{N}} \mu_k \sum_{n=0}^{N-1} x[n] \cos(\frac{\pi}{2N}(2n+1)k),\ \ 0 \le k \le N-1, \tag{2}$$

where $\mu_0 = 1$ and $\mu_k = \sqrt{2}$ for $k > 0$. Besides, $X[k]$ and $C[k]$ are related by

$$\left\{ \begin{array}{cc} X[k] = 2e^{j\frac{\pi k}{2N}} C[k] & ,\ \ 0 \le k \le N-1 \\ X[2N-k] = 2e^{-j\frac{\pi k}{2N}} C[k] & ,\ \ 0 \le k \le N-1 \end{array} \right. \tag{3}$$

It can be shown that the inverse DFT and DCT are:

$$\textbf{IDFT:}\quad x[n] = \tfrac{1}{N} \sum_{k=0}^{N-1} X[k] e^{j\frac{2\pi kn}{N}}\quad,\ \ 0 \le n \le N-1 \tag{4}$$

and

$$\textbf{IDCT:}\quad c[n] = \tfrac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \mu_k C[k] \cos\left[\tfrac{\pi}{2N}(2n+1)k\right]\quad,\ \ 0 \le n \le N-1. \tag{5}$$

As shown in eq. (1), the DFT $X[k]$ of a real-valued sequence is a complex sequence satisfying the conjugate symmetry condition, $X[k] = X^*[\langle -k \rangle_N]$, and thus about one-half ($\lfloor N/2 \rfloor + 1$) DFT points are in fact redundant. However, in the DCT case $C[k]$ and $x[n]$ are equal in length, and in general $C[k]$ is neither symmetric nor anti-symmetric. Therefore, DCT exhibits higher frequency resolution than DFT. In addition, eq. (3) shows DCT can be performed efficiently via the fast algorithms of DFT.

II.2 Properties of DCT

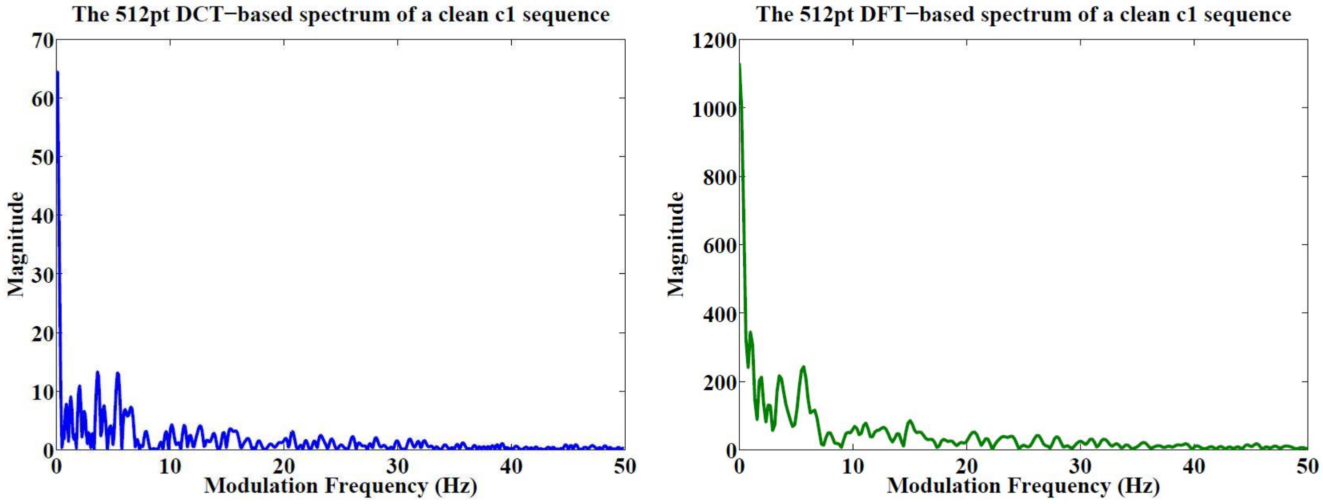
[10] shows the Karhunen Loeve Transform (KLT) gives the optimal performance in transform coding. However, KLT lacks fast algorithms in implementation. DCT compares more closely to KLT in coding performance relative to other orthogonal transforms.Therefore, DCT serves as a very good alternative of KLT for coding speech signals. Besides, DCT provides higher frequency resolution than DFT, and is more efficiently computable than discrete wavelet transform (DWT).

II.3 The impact of noise on the DCT of speech feature stream

When it comes to the analysis for the temporal characteristics of the speech feature stream, we often focus on the DFT-based modulation spectrum. In contrast, the “modulation spectrum” derived from DCT is much less considered. Since DCT possesses peculiar properties relative to DFT as described previously. Here we would like to observe the DCT-based modulation spectrum of a feature stream and investigate the corresponding response to noise.

First, Figures 1(a) and (b) depict the DCT-based and DFT-based modulation (magnitude) spectra for the MFCC c_1 feature stream of a clean utterance. We find that the DCT-based spectrum is more concentrated at low frequencies in energy than the DFT-based spectrum, and it shows higher frequency resolution.

Next, we investigate the impact of noise on the DCT-based modulation spectrum, which is separately observed in magnitude and phase (sign). Note that the DCT of an arbitrary sequence is real-valued, which can be only positive, zero or negative, corresponding a binary phase of 0 and π.

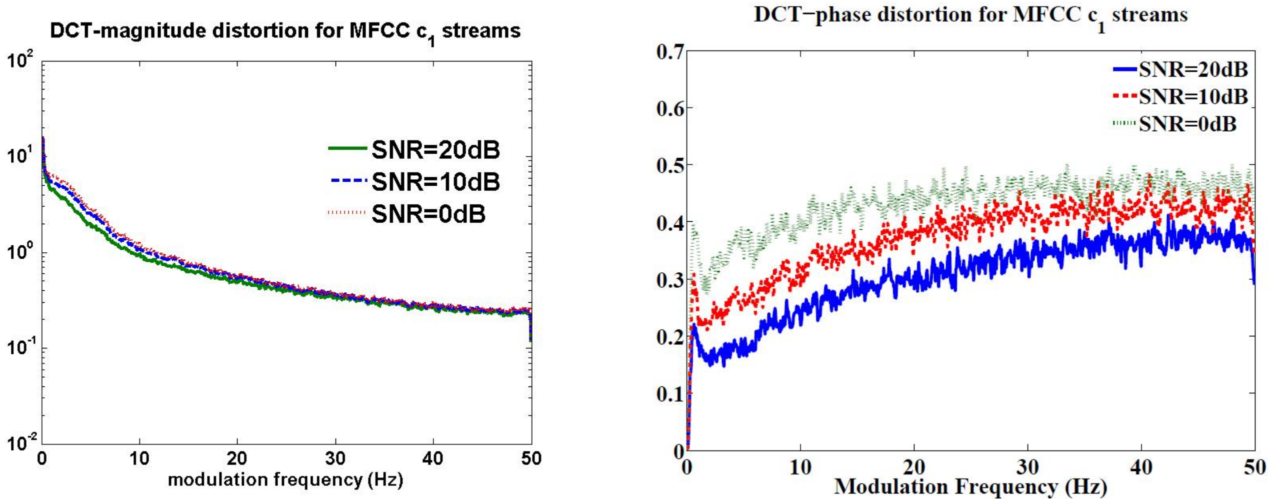


(a)

(b)

Figure 1: The modulation (magnitude) spectrum of (a) DCT-based and (b) DFT-based for the MFCC c_1 feature stream of a clean utterance.

Figures 2(a) and (b) depict the averaged magnitude and phase (sign) distortions by comparing the DCT-based modulation spectra of the MFCC c_1 streams for a set of 1001 clean utterances and its three noisy counterparts at signal-to-noise ratios (SNRs) 20 dB, 10 dB and 0 dB. From Figure 2(a), the DCT-magnitude distortions increase as the SNR get worse, and larger distortion components are mainly located in the low frequency region (roughly [0, 10 Hz]). Besides, Figure 2(b) shows that amplifying the noise level (with a lower SNR) introduces more DCT-phase (sign) distortions. However, in contrast to the case of DCT-magnitudes, DCT-phase distortions are approximately uniformly distributed over the whole frequency range, with the relatively larger phase distortions dwelling at high frequencies probably because the corresponding DCT coefficients are smaller in magnitude and easier to be changed in phase (sign).

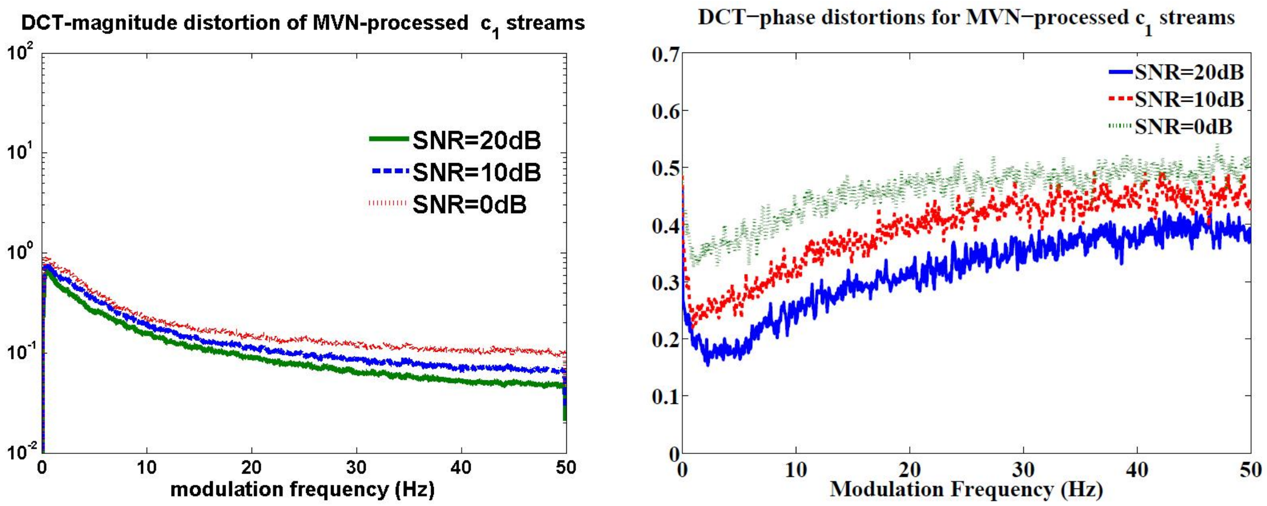


(a)

(b)

Figure 2: The averaged (a) DCT-magnitude distortions and (b) DCT-phase distortions in the original MFCC c_1 streams caused by babble noise at three SNRs, 20 dB, 10 dB and 0 dB.

Moreover, here the well-known noise-robustness method, mean and variance normalization (MVN) [2], is selected to process the MFCC features used in Figures 2(a) and (b), and the corresponding DCT-magnitude and DCT-phase distortions are plotted in Figures 3(a) and (b), respectively. Comparing Figure 3(a) with Figure 2(a), DCT-magnitude distortions are significantly reduced by MVN. On the contrary, DCT-phase distortions shown in Figure 3(b) remain significant as shown in Figure 2(b). These results imply the good performance of MVN mainly comes from its capacity of reducing DCT-magnitude distortions rather than DCT-phase distortions.



(a)

(b)

Figure 3: The averaged (a) DCT-magnitude distortions and (b) DCT-phase distortions in the MVN-processed MFCC c_1 streams caused by babble noise at three SNRs 20 dB, 10 dB and 0 dB.

III. The proposed DCT-based feature compensation approaches

This section is arranged as follows: First, we introduce two new proposed feature compensation methods based on DCT, and they are termed "DCT magnitude substitution" (DCT-MS) and "DCT magnitude weighting" (DCT-MW), respectively. Next, we introduce a variant of DCT-MS, which differs from DCT-MS primarily in the selection of processed frequency range. Finally, we examine these new methods in their capability of reducing the mismatch in the power spectral density (PSD) of feature streams.

III.1 The concepts of DCT-based speech feature compensation methods

According to the discussions in the previous section, the magnitude parts of the DCT for speech feature streams are vulnerable to noise, and properly dealing with them such as the MVN process can help a lot. Here we attempt to provide some directions to alleviate the DCT-magnitude distortions.

Let $\{x[n];\ 0 \leq n \leq L-1\}$ be the temporal-domain feature sequence of an arbitrary utterance for each channel, and its M-point DCT is represented by

$$\{C[k];\ 0 \leq k \leq M-1\}. \tag{6}$$

Then $C[k]$ corresponds to the DCT-based modulation spectrum of $\{x[n]\}$ at frequency $f = k\frac{F_s}{2M}$ in Hz, where F_s (Hz) is the frame rate of $\{x[n]\}$. Note here the DCT-size M is set to be larger than or equal to L, the length of $\{x[n]\}$, to avoid the time aliasing effect. Briefly speaking, our methods update these $C[k]$'s in its magnitude part $|C[k]|$, and leave its sign (phase) part $sgn(C[k])$ unchanged, hoping that the mismatch of $|C[k]|$ among different SNR cases can be thus reduced.

We present two types of DCT-based feature compensation methods, both of which consist of three steps:

Step 1: Obtain the DCT-magnitude reference or the DCT-magnitude weight from the training data:

Let $\{C[k];\ 0 \le k \le M-1\}$ be the M-point DCT of any temporal sequence in *the training set* with respect to a specific channel. Here the used DCT-size M is common to any temporal sequence in the training set, and this setting makes the DCT spectra of all training sequences (with respect to a specific channel) have the same length M. We calculate two sequences:

DCT-magnitude reference:

$$A_{ref}[k] = E\{|C[k]|\} = \frac{1}{N_{ref}} \sum_{C[k]\in training\ set} |C[k]|, \tag{7}$$

and

DCT-magnitude weight:

$$\sigma_{ref}[k] = std\{C[k]\} = \sqrt{\frac{1}{N_{ref}} \sum_{C[k]\in training\ set} C^2[k] - \left(\frac{1}{N_{ref}} \sum_{C[k]\in training\ set} C[k]\right)^2}, \tag{8}$$

where $E\{X\}$ and $std\{X\}$ denote the mean and standard deviation of X, and N_{ref} is the number of $C[k]$'s in the training set.

Step 2: Update the DCT magnitude component of the speech features currently processed:

In Step 1, the DCT-magnitude reference/weight shown in eqs. (7) and (8) are obtained from the feature sequences of **all** the clean utterances in the training set. Now we apply them to update the DCT-magnitude of **each** feature sequence in both the training and testing sets. Briefly speaking, the DCT coefficients $\{C[k];\ 0 \le k \le M-1\}$ of any feature sequence in the training and testing sets is updated in magnitude, and we produce the new DCT stream:

$$\tilde{C}[k] = \left|\tilde{C}[k]\right| sgn(C[k]), \qquad 0 \le k \le M-1. \tag{9}$$

where $|\tilde{C}[k]|$ denotes the new DCT-magnitude. That is, the original and updated DCT streams differ only in magnitude, not in phase. We propose various ways to update the DCT-magnitude, and they will be described in detail in the next subsections.

Step 3: Use IDCT to obtain the new feature sequence:

The the L-point new feature stream is obtained by

$$\tilde{x}[n] = IDCT_M\{\tilde{C}[k];\ 0 \leq k \leq M-1\}, \qquad 0 \leq n \leq L-1. \tag{10}$$

That is, the M-point inverse DCT is performed on the M-point sequence $\{\tilde{C}[k]\}$, and the resulting M-point sequence $\{\tilde{x}[n]\}$ is *truncated* and thus only the first L points in $\{\tilde{x}[n]\}$ are reserved.

III.2 The DCT-magnitude updated algorithms

In this subsection, we provide two different directions to update the DCT-magnitude of a speech feature stream as mentioned in Step 2 of sub-section III.1.

III.2.1 DCT-magnitude substitution (DCT-MS)

In DCT-MS, the DCT-magnitude of each feature stream currently processed is directly substituted by the DCT-magnitude reference shown in eq. (7). That is,

$$|\tilde{C}[k]| = A_{ref}[k], \quad 0 \leq k \leq M-1. \tag{11}$$

This operation is primarily motivated by two observations:

1. The DCT-magnitudes among different clean feature sequences look similar to one another. Using the same DCT-magnitude for different feature sequences probably causes a small amount of distortion.

2. Noise affects the DCT-magnitude very significantly, and thus the DCT-magnitude of a noisy feature stream is highly deviated from that of a clean one. Introducing a unified DCT-magnitude completely removes the effect of noise (while probably loses some speech information).

III.2.2 DCT-magnitude weighting (DCT-MW)

In DCT-MW, the DCT magnitude of each feature stream currently processed is directly multiplied by the DCT-**magnitude weight** defined in eq. (8). That is:

$$|\tilde{C}[k]| = |C[k]|\sigma_{ref}[k], \quad 0 \leq k \leq M-1. \tag{12}$$

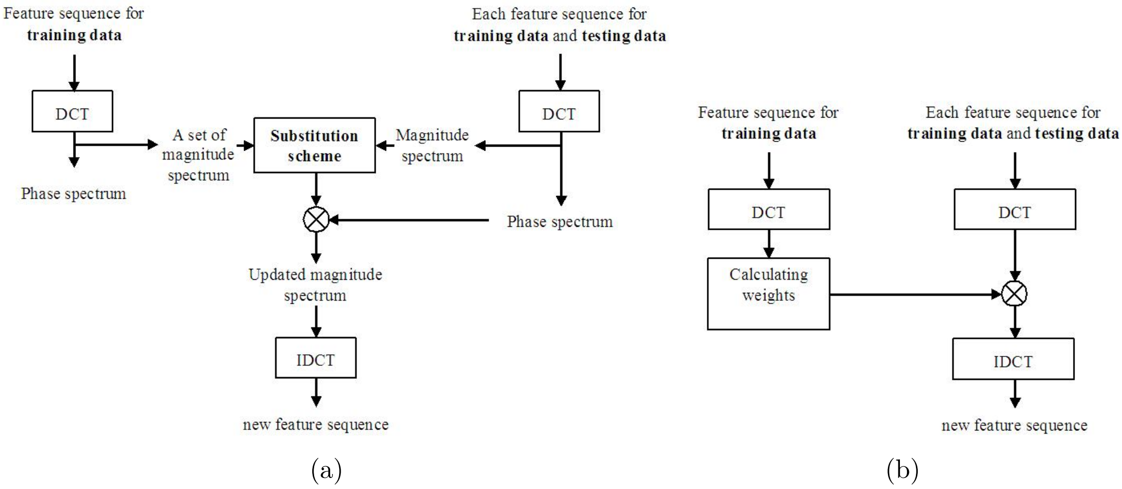


Figure 4: The flowchart of (a)DCT-MS (b)DCT-MW

The method of DCT-MW is basically from two ideas:

1. In general, the variance, or its variant such as the standard deviation, accounts for the amount of gross information contained in a random variable. Assuming most of the information corresponds to speech, to weigh the noisy DCT-magnitude with the standard deviation of the clean DCT-magnitudes probably highlights the speech components.

2. The original noisy DCT-magnitude, that is expected to contain speech information and benefit the recognition, is reserved in DCT-MW. Furthermore, DCT-MW behaves similarly to a zero-phase temporal filter, which can effectively improve the noise robustness of features if properly designed.

The flowcharts of DCT-MS and DCT-MW are depicted in Figures 4(a) and (b). Besides, the DCT-magnitude weight for DCT-MW from the MVN-processed MFCC c_1 streams is plotted in Figure 5, which shows the DCT-magnitudes at lower modulation frequencies are to be amplified in DCT-MW. This is somewhat consistent to the general idea that, the modulation frequency components within [1 Hz, 16 Hz] contain rich speech information [11], and emphasizing these components properly will improve the recognition accuracy.

III.2.3 Partial-band DCT-MS

The substitution process of DCT-MS is originally carried out on the entire DCT-magnitude stream, indicating that each modulation frequency component within the full-band range [0, $\frac{F_s}{2}$ Hz] is updated, where F_s is the frame rate in Hz. Here, we propose to select the components within a specific partial-band rather than the full-band to perform DCT-MS.

This partial-band process is mainly inspired by two considerations:

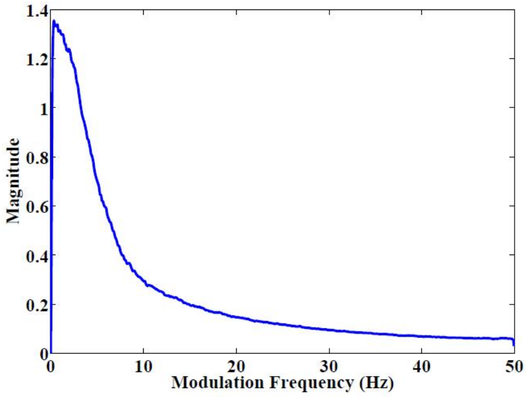


Figure 5: The DCT-magnitude weight for MVN-processed MFCC c_1 features in DCT-MW.

1. Keeping the less-distorted components unchanged:
   The deviations in the DCT-magnitudes caused by noise are in fact unequal. In particular, noise probably just contaminates a few frequency components primarily. Updating the DCT-magnitudes at all frequencies introduces another distortion, especially to those less noise-affected ones.

2. Reducing the computation complexity:
   Provided that the recognition accuracy is not degraded, decreasing the number of DCT-magnitudes necessary for an update reduces the computation complexity of the algorithms for sure.

Here, we arrange the partial-band version of DCT-MS by simply setting a cutoff frequency F_c, dividing the frequency range into two sub-bands [0, F_c Hz] and [F_c Hz, $\frac{F_s}{2}$ Hz], and performing DCT-MS for either one sub-band. Accordingly, the performance of the patial-band DCT-MS depends on the selection of the cutoff frequency F_c and the sub-band components to be updated.

Note that we do not provide the partial-band version of DCT-MW since it seems not very appropriate to weigh some DCT-magnitudes and leave the others unchanged, which behaves like a filter having a discontinuity at the cutoff frequency in magnitude response.

III.3 A preliminary evaluation of DCT-MS/DCT-MW in reducing the noise effect

We perform the proposed DCT-MS or DCT-MW on the MFCC c_1 feature streams of three utterances containing the same embedded clean speech while distorted at different SNRs: clean, 10 dB and 0 dB with subway noise. Before acting DCT-MS/DCT-MW, the feature sequence is processed by MVN to be zero-mean and unity-variance.

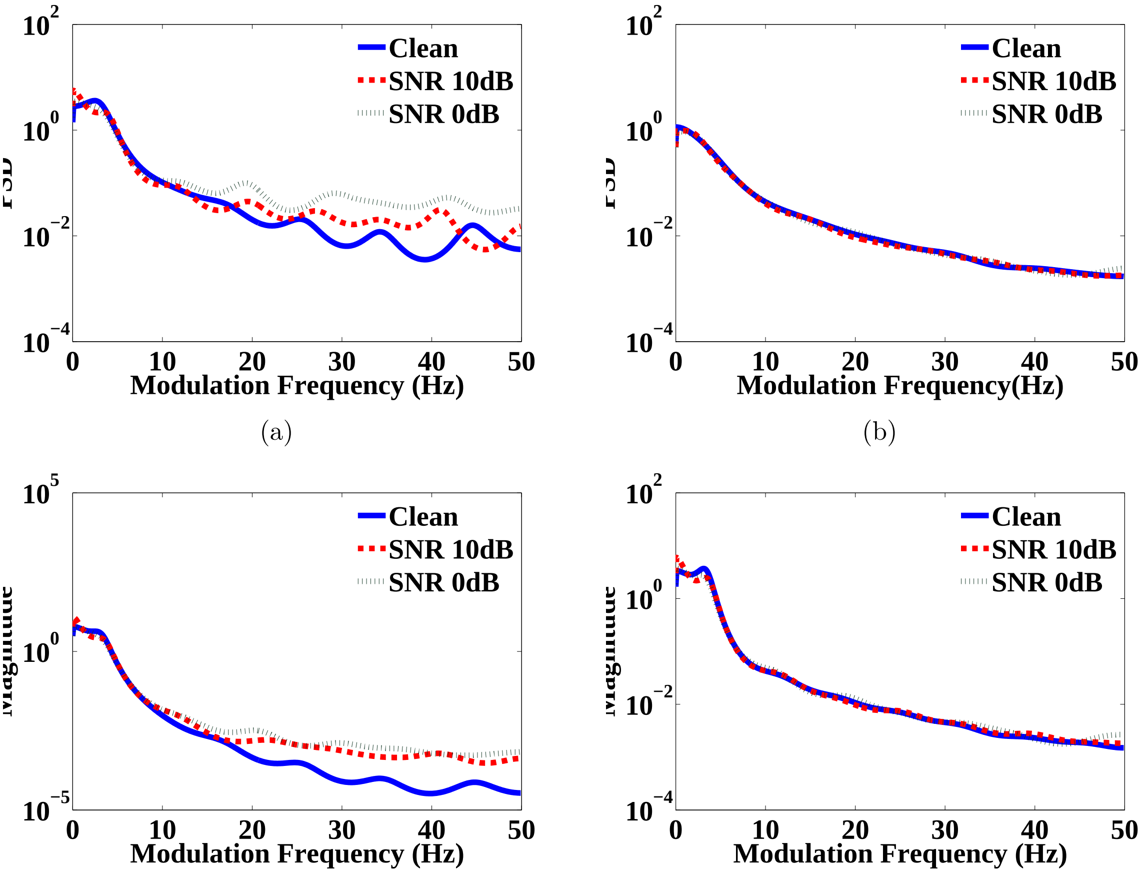
Figures 6(a)-(d) plot the power spectral density (PSD) curves of the c_1 feature streams for three SNR cases obtained from various processes. The corresponding detailed information and discussions are:

1. As shown in Figure 6(a), there exists significant mismatch among the PSDs of original (MVN-processed) features at different SNRs. The mismatch gets larger with increasing frequency. The PSD becomes relatively ”flat” as the SNR gets worse, which agrees with the observation in [8].

2. Figure 6(b) corresponding to the features processed by DCT-MS reveals that this method successfully reduces the PSD mismatch shown in Figure 6(a). The direct substitution for the DCT-magnitudes of different feature streams with a common reference curve makes the associated PSD curves so close to each other.

3. From Figure 6(c), the PSDs of DCT-MW processed features still contain significantly mismatch as the ones from MVN in Figure 6(a). However, the scale of deviation (for the frequency greater than 10 Hz) shown in Figure 6(c) is below 10^{-2}, while the original PSD deviation shown in Figure 6(a) is roughly within the range $[10^{-1}, 10^{-2}]$. As a result, DCT-MW can reduce the PSD mismatch effectively.

4. Figure 6(d) depicts the PSDs for the “partial-band” version of DCT-MS, in which the frequency range to be updated is set to [5 Hz, 50 Hz]. That is, the first one-tenth band [0, 5 Hz] components are kept unchanged. We find that they are quite similar to the curves shown in Figure 6(b) (the “full-band” version of DCT-MS): the median/high frequency distortion is insignificant. The unprocessed band [0, 5 Hz] appears deviations among the curves. The positive or negative effect of keeping the low frequency components unchanged in recognition accuracy will be shown in section IV.



(a)

(b)

(c)

(d)

Figure 6: The c_1 PSD curves processed by various methods:(a)MVN (b)DCT-MS (c)DCT-MW (d)partial-band DCT-MS

IV. The recognition experiment results and discussions

This section is organized as follows: Firstly, sub-section IV.1 introduces the used speech database and the setup for the experimental environment. Secondly, the recognition results for the original and MVN-processed MFCC are provided in sub-section IV.2. Thirdly, we present and discuss the recognition accuracy obtained by the new DCT-based algorithms in sub-section IV.3. Finally, sub-section IV.4 briefly summarizes the recognition results of the DCT-based algorithms for the features preliminary processed by some robustness methods.

IV.1 The Experimental Environmental Setup

Our recognition experiments are conducted on the AURORA 2.0 database , the details of which are described in [12]. In short, the testing data consist of 4004 utterances from 52 female and 52 male speakers, and three different subsets are defined for the recognition experiments: Test Sets A and B are each affected by four types of noise, and Set C is affected by two types.

Each noise instance is added to the clean speech signal at six SNR levels (ranging from 20 dB to -5 dB). The signals in Test Sets A and B are filtered with a G.712 filter, and those in Set C are filtered with an MIRS filter. In the "clean-condition training, multi-condition testing" mode defined in [12], the training data consist of 8440 *clean* speech utterances from 55 female and 55 male adults. These signals are filtered with a G.712 filter. The data in Test Sets A and B are more distorted by additive noise than the training data, while the data in Set C are affected by additive noise and a channel mismatch.

With the Aurora-2 database, we performed the a series of robustness methods to compare the recognition accuracy. Each utterance in the clean training set and three testing sets is directly converted to 13-dimensional MFCC ($c0 \sim c12$) sequence. Next, the MFCC features are then updated by either noise-robustness method. The resulting 13 new features, plus their first- and second-order derivatives, are the components of the final 39-dimensional feature vector. With the new feature vectors in the clean training set, the hidden Markov models (HMMs) for each digit and silence are trained with the HTK toolkit [13] . Each digit HMM has 16 states, with 20 Gaussian mixtures per state.

IV.2 Experiment results of plain MFCCs and MVN-processed MFCCs

The recognition accuracy rates for the original MFCC are shown in Table 1. From this table, we have some observations as follows:

1. When the SNR becomes worse, the recognition accuracy rate gets lower in every noisy environment. Therefore, noise brings a significant distortion to MFCC features.

2. The averaged recognition accuracy of Set A is better than that of Set B probably because most noise types in Set A are stationary and most noise types in Set B are non-stationary.

3. Among the four noise types in Set A, "babble" and "exhibition" result in the largest and smallest accuracy degradation, respectively. In contrast, the noise types in Set B that correspond to the highest and lowest accuracy rates are "airport" and "street".

4. With the same noise type "subway", the accuracy of Set A is better than that of Set C, implying the channel mismatch in Set C further degrades the recognition performance.

Among the various noise-robustness algorithms,MVN is very widely used since implementing MVN is quite simple and significant recognition improvement can be thus achieved. Many noise-robustness techniques such as TSN [7] and MVA [8] have been developed directly on MVN-processed MFCC features and reveals very good performance. As a result, we treat the MVN-processed MFCC as the baseline features hereafter, unless otherwise mentioned.

The recognition results of the baseline experiments, using MVN-processed MFCC as features, are shown in Table 2. Comparing Table 2 with Table 1, MVN benefits the plain MFCC a lot

Table 1: The recognition accuracy rates (%) of plain MFCCs in various environments

baseline experiments
(using MFCCs, including $c_0 \sim c_{12}$ plus their delta and delta-delta, totally 39 features)

	Set A					Set B					Set C		
	subway	babble	car	exhibition	average	restaurant	street	airport	train	average	subway	street	average
clean	99.83	99.77	99.74	99.85	99.80	99.83	99.77	99.74	99.85	99.80	99.79	99.76	99.78
20dB	98.90	91.26	97.14	98.72	96.51	94.41	97.33	92.87	93.67	94.57	96.58	97.16	96.87
15dB	95.08	78.27	88.16	95.25	89.19	84.12	92.22	80.54	83.06	84.99	91.63	93.16	92.40
10dB	82.43	61.68	69.65	84.04	74.45	67.83	77.61	63.88	66.07	68.85	82.24	82.64	82.44
5dB	62.31	44.41	53.20	63.63	55.89	49.55	60.19	48.38	49.28	51.85	65.01	67.25	66.13
0dB	47.12	33.20	45.00	49.04	43.59	36.13	47.74	37.98	41.52	40.84	48.64	51.79	50.22
-5dB	43.13	30.89	42.60	43.77	40.10	33.60	42.81	35.42	40.15	38.00	43.58	45.33	44.46
average	77.17	61.76	70.63	78.14	**71.92**	66.41	75.02	64.73	66.72	**68.22**	76.82	78.40	**77.61**

Table 2: The recognition accuracy rates (%) of the baseline experiment, with the MVN-processed MFCC as the features

Baseline experiment results (with MVN-processed MFCC features)

	Set A					Set B					Set C		
	subway	babble	car	exhibition	average	restaurant	street	airport	train	average	subway	street	average
clean	99.81	99.77	99.76	99.92	99.82	99.81	99.77	99.76	99.92	99.82	99.85	99.79	99.82
20dB	98.46	99.06	98.71	98.32	98.64	99.20	98.72	99.12	98.47	98.88	98.52	98.74	98.63
15dB	96.73	96.95	96.73	96.22	96.66	97.62	96.82	97.67	96.05	97.04	96.79	96.76	96.78
10dB	92.03	92.20	90.91	90.90	91.51	93.34	91.54	93.24	91.05	92.29	91.92	91.64	91.78
5dB	81.25	78.68	74.90	81.08	78.98	81.95	79.10	80.63	76.96	79.66	81.21	79.47	80.34
0dB	62.39	57.61	53.56	63.89	59.36	63.55	59.11	61.31	55.66	59.91	61.97	58.96	60.47
-5dB	47.84	45.63	43.72	48.64	46.46	48.17	46.44	46.98	45.30	46.72	47.58	46.74	47.16
average	86.17	84.90	82.96	86.08	**85.03**	87.13	85.06	86.39	83.64	**85.56**	86.08	85.11	**85.60**
MFCC	77.17	61.76	70.63	78.14	71.92	66.41	75.02	64.73	66.72	68.22	76.82	78.40	77.61

by enhancing the recognition accuracy rates for almost all SNR cases and all noise types, which exhibits the capability of improving noise robustness of MVN for MFCC. Furthermore, even though MVN does not eliminate the median/high (modulation) frequency distortion very well, as depicted in Figure 3(a), the low-frequency portion that contains most speech information is well treated by MVN in reducing noise effects, thus bringing about very good recognition accuracy.

IV.3 The experimental results of proposed DCT-based algorithms

IV.3.1 DCT-MS and DCT-MW

This sub-section provides the results of DCT-MS and DCT-MW. The parameter M in eq. (6) that represents the length of the common DCT-magnitude reference/weight for DCT-MS/ DCT-MW is set to 1024.

Tables 3 and 4 give the detailed recognition accuracy rates obtained from DCT-MS and DCT-MW. We have some findings from the two tables:

1. Compared with the baseline results in Table 2, both DCT-MS and DCT-MW provide better recognition accuracy, implying the two methods can enhance MVN features in noise robustness.
2. DCT-MW outperforms DCT-MS slightly,indicating that to highlight the more important DCT-components like a filtering process helps more. For example, with DCT MW, the averaged accuracy for Set B can be as high as 90%, roughly 4% better than the baseline.

Table 3: The recognition accuracy rates (%) of DCT-MS that performs on the MVN-processed MFCC

DCT-MS													
	Set A					**Set B**					**Set C**		
	subway	**babble**	**car**	**exhibition**	**average**	**restaurant**	**street**	**airport**	**train**	**average**	**subway**	**street**	**average**
clean	99.37	99.23	99.25	99.58	99.36	99.37	99.23	99.25	99.58	99.36	99.43	99.11	99.27
20dB	97.91	98.38	98.73	98.13	98.29	98.35	98.23	98.62	98.63	98.46	98.14	98.32	98.23
15dB	96.08	96.93	97.55	96.43	96.75	97.39	97.21	97.87	97.50	97.49	96.48	96.89	96.69
10dB	92.34	94.12	94.38	92.68	93.38	94.09	93.92	95.39	94.70	94.53	92.09	93.63	92.86
5dB	84.08	85.97	87.86	85.18	85.77	86.73	87.08	88.31	88.59	87.68	84.52	87.42	85.97
0dB	71.10	69.64	76.34	71.98	72.27	72.68	74.44	75.69	75.62	74.61	70.55	74.80	72.68
-5dB	56.34	52.56	61.46	57.24	56.90	55.20	59.04	58.26	60.37	58.22	56.08	59.17	57.63
average	88.30	89.01	90.97	88.88	**89.29**	89.85	90.18	91.18	91.01	**90.55**	88.36	90.21	**89.28**
MVN baseline	86.17	84.90	82.96	86.08	85.03	87.13	85.06	86.39	83.64	85.56	86.08	85.11	85.60

IV.3.2 Partial-band DCT-MS

Here we perform the partial-band DCT-MS given in sub-section III.2.3. For the sake of clarity, the notations $_p$DCT-MS$_u$ and $_p$DCT-MS$_l$ are used, where the left subscript index "p" indicates a *p*artial-band DCT-MS, and the right subscript, "u" or "l", represents the updated partial band being "*u*pper sub-band" ([F_c Hz, $F_s/2$ Hz]) or "*l*ower sub-band" ($[0, F_c$ Hz]), in which F_c and F_s are the cutoff frequency and the frame rates in Hz. The averaged recognition accuracy rates achieved by $_p$DCT-MS$_u$ and $_p$DCT-MS$_l$ for five different assignments of cutoff frequency F_c are listed in Tables 5 and 6. We have the following observations from the two tables:

1. For the case of $_p$DCT-MS$_u$, in which only the upper sub-band magnitudes are updated and increasing the cutoff frequency narrows the upper sub-band in bandwidth, the corresponding recognition accuracy rates are always better than the baseline (with MVN-processed

Table 4: The recognition accuracy rates (%) of DCT-MW that performs on the MVN-processed MFCC

DCT-MW													
	Set A					**Set B**					**Set C**		
	subway	**babble**	**car**	**exhibition**	**average**	**restaurant**	**street**	**airport**	**train**	**average**	**subway**	**street**	**average**
clean	99.66	99.53	99.66	99.83	99.67	99.66	99.53	99.66	99.83	99.67	99.64	99.57	99.61
20dB	98.75	98.95	98.95	98.55	98.80	99.20	98.55	98.94	98.91	98.90	98.76	98.53	98.65
15dB	97.76	97.53	97.61	96.47	97.34	98.21	97.72	98.11	97.52	97.89	97.43	97.66	97.55
10dB	94.20	94.12	95.13	92.47	93.98	94.92	94.78	95.13	94.72	94.89	93.90	94.76	94.33
5dB	86.31	85.29	88.40	84.40	86.10	86.37	87.14	87.64	87.74	87.22	86.31	87.16	86.74
0dB	70.26	66.50	74.75	71.58	70.77	68.66	72.88	72.25	72.57	71.59	70.22	72.11	71.17
-5dB	53.68	48.87	56.60	55.90	53.76	50.47	54.16	54.06	55.39	53.52	53.26	54.01	53.64
average	89.46	88.48	90.97	88.69	**89.40**	89.47	90.21	90.41	90.29	**90.10**	89.32	90.04	**89.68**
MVN baseline	86.17	84.90	82.96	86.08	85.03	87.13	85.06	86.39	83.64	85.56	86.08	85.11	85.60

features). However, ${}_p$DCT-MS$_u$ outperforms the full-band DCT-MS (with the cutoff frequency 0 Hz) only when the cutoff frequency F_c is 5 Hz, and there is a performance gap when F_c is from 5 Hz to 15 Hz. This observation leads to two aspects: First, keeping the components within [0, 5 Hz] unchanged is better than updating them, probably because this frequency range has been handled well by MVN and further normalizing it in DCT-magnitude tends to attenuate the recognition information. Second, operating DCT-MS in the frequency range [5 Hz, 15 Hz] especially helps in recognition performance, which is somewhat consistent of the observation in Figure 3(a) that there remains PSD mismatch roughly above 5 Hz after operating MVN.

2. For the case of ${}_p$DCT-MS$_l$, in which only the lower sub-band magnitudes are updated and increasing the cutoff frequency broadens the lower sub-band in bandwidth, assigning a too small cutoff frequency (below 10 Hz) even worsens the recognition accuracy relative to the baseline, which supports our statements for ${}_p$DCT-MS$_u$ previously that updating the components within the frequency range [0, 5 Hz] is not a good idea. Increasing the cutoff frequency F_c in ${}_p$DCT-MS$_l$ can improve the recognition accuracy, and the best possible results for ${}_p$DCT-MS$_l$ occurs when F_c is 50 Hz, equivalent to the original (full-band) DCT-MS. As a result, partial-band DCT-MS outperforms full-band DCT-MS only when a proper *upper* sub-band is selected for update.

IV.3.3 Integrating DCT-MS/DCT-MW with other normalization techniques

In sub-section IV.3.2 the MVN-processed MFCC are treated as the baseline features and they are further updated with the presented DCT-based algorithms. Experimental results show that the DCT-based algorithms achieve higher recognition accuracy relative to the baseline,

Table 5: Recognition accuracy rates (%) averaged over all noise types different SNRs for the $_{p}$DCT-MS$_{u}$ method with different cutoff frequency, where AR(%) and RR(%) stand for the absolute and relative error rate reductions, respectively.

$_{p}$**DCT-MS**$_{u}$ **(updating the upper sub-band) with different cutoff frequencies**						
Cutoff frequency F_c	**Set A**	**Set B**	**Set C**	**Average**	**AR**	**RR**
0 Hz (full-band DCT-MS)	89.29	90.55	89.28	89.79	4.44	30.31
5 Hz	90.80	91.62	90.12	90.99	5.64	38.50
15 Hz	87.51	88.03	87.95	87.80	2.45	16.72
25 Hz	86.04	86.60	86.63	86.38	1.03	7.03
35 Hz	85.57	86.14	86.24	85.93	0.58	3.96
45 Hz	85.16	85.78	85.72	85.52	0.17	1.16
50 Hz(equivalent to the baseline)	85.03	85.56	85.60	85.35	-	-

Table 6: Recognition accuracy rates (%) averaged over all noise types different SNRs for the $_{p}$DCT-MS$_{l}$, with different cutoff frequency, where AR(%) and RR(%) stand for the absolute and relative error rate reductions, respectively.

$_{p}$**DCT-MS**$_{l}$ **(updating the lower sub-band) with different cutoff frequencies**						
Cutoff frequency F_c	**Set A**	**Set B**	**Set C**	**Average**	**AR**	**RR**
50 Hz(full-band DCT-MS)	89.29	90.55	89.28	89.79	4.44	30.31
45 Hz	89.13	90.50	89.16	89.68	4.33	29.56
35 Hz	88.59	89.98	88.75	89.18	3.83	26.14
25 Hz	88.27	89.70	88.46	88.88	3.53	24.10
15 Hz	85.73	87.26	86.05	86.41	1.06	7.24
5 Hz	83.78	84.71	84.53	84.30	-1.05	-7.17
0 Hz(equivalent to the baseline)	85.03	85.56	85.60	85.35	-	-

revealing that they are well additive to MVN. Here we are to investigate if the proposed DCT-MS/DCT-MW can enhance some other types of features, including the original plain MFCCs and the MFCCs processed by either of CMN, CGN, MVA, and HEQ in advance.

Tables 7, 8 and 9 list the averaged recognition accuracy rates for DCT-MS, DCT-MW and $_{p}$DCT-MS$_{u}$ ($F_c = 5$ Hz), respectively, for different types of features (MFCCs processed by CMN, MVN, CGN, HEQ and MVA). From the three tables, we find that

1. Similar to MVN, all the pre-processing algorithms including CMN, CGN, HEQ and MVA provide the original MFCC with improved recognition accuracy. MVA performs the best, followed by HEQ, CGN, MVN and then CMN.

2. The presented DCT-MS enhances the recognition accuracy for all the types of features shown here, including the unprocessed plain MFCCs. The resulting average accuracy rates are around 89.50% (except DCT-MS performing on the plain MFCCs). As a result,

Table 7: Recognition accuracy rates (%) averaged over all noise types different SNRs for the DCT-MS method combined with various featuer normalization methods

DCT-MS on various feature normalization methods						
Method	**Set A**	**Set B**	**Set C**	**Average**	**AR**	**RR**
MFCC	71.92	68.22	77.61	71.58	-	-
MFCC+DCT-MS	82.73	84.55	83.39	83.59	12.01	42.26
CMN	79.37	82.47	79.90	80.71	-	-
CMN+DCT-MS	89.15	90.45	89.23	89.68	8.97	46.50
MVN	85.03	85.56	85.60	85.35	-	-
MVN+DCT-MS	89.29	90.55	89.28	89.79	4.44	30.31
HEQ	87.59	88.84	87.64	88.10	-	-
HEQ+DCT-MS	88.50	90.00	89.04	89.21	1.11	9.33
CGN	87.64	88.55	87.73	88.02	-	-
CGN+DCT-MS	89.25	90.58	89.27	89.79	1.77	14.77
MVA	88.12	88.81	88.50	88.47	-	-
MVA+DCT-MS	88.93	90.20	88.88	89.42	0.95	8.24

by adopting DCT-MS, CMN and CGN become more attractive than HEQ and MVA since they are more computationally efficient.

3. Similar to DCT-MS, integrating DCT-MW with most normalization methods (except CMN and the original MFCC) provide better recognition rates than the individual component method. The optimal performance, 90.84% in averaged accuracy, occurs with the pairing of DCT-MW and CGN, better than those shown in Table 8, indicating DCT-MW behaves better than DCT-MS when combining with any of CGN, HEQ and MVA. However, since there remains significant low modulation frequency distortion in the unprocessed and CMN-processed noisy MFCC features, DCT-MW, acting as a low-pass filter, cannot benefit the two types of features in reducing the effect of noise.

4. Similar to DCT-MS and DCT-MW, ${}_p$DCT-MS$_u$ (with $F_c = 5$ Hz) is well additive to most normalization methods to make the recognition accuracy better. Comparing Table 9 with Tables 7 and 8, the partial-band DCT-MS, ${}_p$DCT-MS$_u$, outperforms the full-band DCT-MS and DCT-MW in most cases. The optimal averaged recognition accuracy shown in Table 9 is as high as 91.41%, with the pairing of ${}_p$DCT-MS$_u$ and HEQ.

IV.4 Summary

The averaged recognition accuracy rates for some methods presented in sub-section IV.3 are summarized in Figure 7 for a clear comparison. From this figure, we find that: First, among the three DCT-based algorithms, only DCT-MS can enhance the original and CMN-processed MFCC features to achieve a high accuracy rate as 89%. Second, when integrating either MVN,

Table 8: Recognition accuracy rates (%) averaged over all noise types different SNRs for the DCT-MW method combined with various featuer normalization methods

DCT-MW on various feature normalization methods						
Method	**Set A**	**Set B**	**Set C**	**Average**	**AR**	**RR**
MFCC	71.92	68.22	77.61	71.58	-	-
MFCC+DCT-MW	74.28	74.44	68.03	73.09	1.51	5.31
CMN	79.37	82.47	79.90	80.71	-	-
CMN+DCT-$MW_{(1)}$	80.02	83.05	80.60	81.35	0.64	3.32
MVN	85.03	85.56	85.60	85.35	-	-
MVN+$MW_{(1)}$	89.40	90.10	89.68	89.74	4.39	29.97
HEQ	87.59	88.84	87.64	88.10	-	-
HEQ+DCT-MW	90.24	90.80	90.85	90.59	2.49	20.92
CGN	87.64	88.55	87.73	88.02	-	-
CGN+DCT-MW	90.39	91.34	90.73	90.84	2.82	23.54
MVA	88.12	88.81	88.50	88.47	-	-
MVA+DCT-MW	89.83	90.59	90.22	90.21	1.47	15.09

Table 9: Recognition accuracy rates (%) averaged over all noise types different SNRs for the $_p$DCT-MS$_u$ method (with $F_c = 5$ Hz) combined with various featuer normalization methods

$_p$DCT-MS$_u$ on various feature normalization methods						
Method	**Set A**	**Set B**	**Set C**	**Average**	**AR**	**RR**
MFCC	71.92	68.22	77.61	71.58	-	-
MFCC+$_p$DCT-MS$_u$	70.33	68.20	75.64	70.54	-1.04	-3.66
CMN	79.37	82.47	79.90	80.71	-	-
CMN+$_p$DCT-MS$_u$	82.69	85.18	83.24	83.79	3.08	15.97
MVN	85.03	85.56	85.60	85.35	-	-
MVN+$_p$DCT-MS$_u$	90.80	91.62	90.12	90.99	5.64	38.50
HEQ	87.59	88.84	87.64	88.10	-	-
HEQ+$_p$DCT-MS$_u$	91.14	92.06	90.66	91.41	3.31	27.82
CGN	87.64	88.55	87.73	88.02	-	-
CGN+$_p$DCT-MS$_u$	90.97	91.87	90.31	91.20	3.18	26.54
MVA	88.12	88.81	88.50	88.47	-	-
MVA+$_p$DCT-MS$_u$	90.45	91.32	90.20	90.75	2.28	19.77

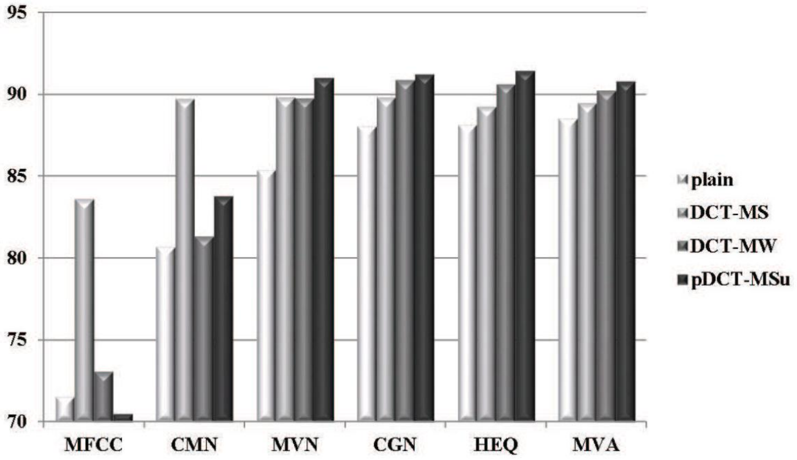


Figure 7: The recognition rates (%) averaged over all noise types and all SNRs for various DCT-based algorithms performing on various types of features

CGN, HEQ or MVA, the partial-band DCT-MS, ${}_p$DCT-MS$_u$, behaves the best, followed by DCT-MW and then DCT-MS. Finally, a relatively computationally efficient algorithm which integrates ${}_p$DCT-MS$_u$ and MVN/CGN can achieve nearly optimal recognition performance since ${}_p$DCT-MS$_u$ is the simplest among the DCT-based algorithms in implementation, and MVN and CGN need less computation complexity than MVA and HEQ.

V. Conclusion and Future Work

In this paper, we use the DCT to develop algorithms to promote the noise robustness of speech features in the temporal domain. In the presented methods, the DCT-magnitudes of feature streams are either normalized or weighted appropriately according to the information of clean speech utterances. We have shown that the two methods give rise to significant word error rate reduction when performing on the MVN-processed features, and they are also well additive to each of CMN, CGN, HEQ and MVA to provide further improved accuracy rates relative to the individual component method.

The future work will be along the following directions:

1. Performing DCT-magnitude substitution adaptively: in this paper we process the DCT-magnitude substitution by directly referring to a fixed reference magnitude curve. Although it may be the most direct and simplest approach, doing this way probably loses some important information of the original noisy speech streams for the ASR task. Therefore, we will study how to collect the information of the currently processed feature stream in order to create the reference magnitude curve in an adaptive manner.

2. Integrating the proposed new methods with some other feature normalization techniques,

such as HOCMN [6] and CSN [4], to see if further improvement can be achieved.

3. Investigating how to determine the optimal trade-off between the noise reduction and the speech distortion that always exists among the noise-robustness techniques.

References

[1] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 254-272, 1981.

[2] O. Viikki and K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication*, vol. 25, pp. 133-147, 1998.

[3] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, “Cpestral Gain Normalization for Noise Robust Speech Recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 209-212, 2004.

[4] Jun Du and Ren-Hua Wang, “Cepstral Shape Normalization (CSN) for Robust Speech Recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4389-4392, 2008.

[5] Ángel de la Torre, Antonio M. Peinado, José C. Segura, José L. Pérez-Córdoba, Ma Carmen Benítez, Antonio J. Rubio, “Histogram Equalization of Speech Representation for Robust Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 355-366, 2005.

[6] C. Hsu and L. Lee, “Higher order cepstral moment normalization (HOCMN) for robust speech recognition,” *Internation Conference on Acoustics, Speech and Signal Processing*, pp. 197-200, 2004.

[7] Xiong Xiao, Eng Siong Chng and Haizhou Li, “Normalization of the Speech Modulation Spectra for Robust Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1662-1674, 2008.

[8] C. Chen and J. Bilmes, “MVA processing of speech features,” *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 257-270, 2006.

[9] S. A. Khayam, “The discrete cosine transform (DCT): theory and application,” *Technical Report WAVES-TR-ECE802.602*, 2003.

[10] Rao, K. and Ahmed, N., “Orthogonal transforms for digital signal processing,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.1, pp. 136-140, 1976.

[11] Noboru Kanedera, Hynek Hermansky and Takayuki Arai, “On properties of modulation spectrum for robust automatic speech recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 613-616, 1998.

[12] H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition system under noisy conditions,” *Proceedings of ISCA IIWR ASR2000*, 2000.

[13] The hidden Markov model toolkit. Available from: <http://htk.eng.cam.ac.uk>.