Robust MFCCs Derived from Differentiated Power Spectrum

J. Chen†, K. K. Paliwal†*, M. Mizumachi* and S. Nakamura**
* ATR Spoken Language Translation Research Laboratories
Kyoto, 619-0288, Japan
† School of Microelectronic Engineering, Griffith University
Brisbane, QLD 4111, Australia
E-mail: jingdong.chen@slt.atr.co.jp

Abstract

The mel-scaled frequency cepstral coefficients (MFCCs) derived from Fourier transform and filter bank analysis are perhaps the most widely used front-ends in state-of-the-art speech recognition systems. One of the major issues with the MFCCs is that they are very sensitive to additive noise. To improve the robustness of speech front-ends with respect to noise, we introduce, in this paper, a new set of MFCC vector which is estimated through three steps. First, the power spectrum of speech signal is estimated through the fast Fourier transform (FFT). Then the power spectrum is differentiated with respected to frequency. Finally, the differentiated power spectrum is transformed into MFCC-like coefficients. Speech recognition experiments for various tasks indicate that the new feature vector is more robust than traditional mel-scaled frequency cepstral coefficients (MFCCs) in additive noise conditions.

1. Introduction

Speech signal carries information from many sources. But not all information is relevant or important for speech recognition. In speech recognition, the first step often called feature analysis or front-end processing is designed to convert the speech signal into some acoustic features which hopefully only encapsulate the important information that is necessary for recognition. Once these features are computed, a post-end classifier is used to classify the input speech signal into a sequence of words in light of the extracted feature vectors and pre-trained models.

Acoustic features may greatly affect the performance of a speech recognizer. Ideally, the selected features should have the capability to discriminate among different acoustic units and, at the same time, be robust with respect to various distortions such as additive noise and channel effects.

A great deal of work has been done for feature selection [1]. The mel-scaled frequency cepstral coefficients (MFCCs) derived though Fourier transform and from filter bank analysis are perhaps the most commonly used acoustic features in currently available speech recognition systems. Much evidence has shown that the MFCCs have served as very successful front-ends for hidden Markov model (HMM) based speech recognition in the past decade. Many speech recognition systems based on these front-ends have achieved very high level of accuracy in clean speech environment.

However, the MFCCs are also found to be sensitive to the noise distortion, especially the additive noise. To improve the robustness of a speech recognition system with respect to noise, in the literature, various methods have been proposed such as Wiener filtering [3], Kalman filtering [4], spectral subtraction [5], RASTA [6], Cepstral mean removal [7], signal bias removal [8], Parallel model compensation (PMC) [10], vector Taylor series approximation based model compensation [11], Jacobian approach [9][12], MLLR [13], transfer vector interpolation [14], *etc*. These methods often take advantage of the prior knowledge of noise to mask, cancel or remove the noise during front-end processing or adjust the system parameters to match the new noisy environment to improve the system performance. There are also some papers focusing on extracting noise resistant feature set such as sub-band based features [18], perception inspired features [19], etc.

Although the aforementioned efforts were experimented in speech recognition with certain success, there remains a great need to investigate new techniques that can accurately recognize speech in degraded environments.

In this paper, we will investigate the use of differentiated power spectrum (DPS) for speech recognition. First, the power spectrum of speech signal is estimated through the FFT. Then the estimated power spectrum is differentiated with respected to frequency. Next, an absolute operation is applied to the DPS. Finally, the magnitude of the DPS is converted to some MFCC-like features by passing it through a mel-scaled frequency filter bank whose output is followed by a log operation and a DCT.

Speech recognition experiments for various tasks in different noise conditions indicate that the proposed approach results in a set of features which are much more robust with respect to noise as compared with the traditional MFCCs.

2. Differential Power Spectrum

2.1 Definition of the Differential Power Spectrum

If $s(t)$ is the original clean speech signal, the received speech signal $y(t)$ is modeled as

$$y(t) = s(t) * h(t) + n(t) = x(t) + n(t) \tag{1}$$

where $h(t)$ is the impulse response of channel distortion and $n(t)$ the ambient noise. $*$ denotes the convolution operation, and $x(t)$ the noise-free speech.

Speech signal is time-variant and non-stationary. It is usually analyzed on the frame-by-frame basis. If we assume that $y_i(n)$ ($1 \le n \le N$, where N is the frame length) represents the *i*th frame of a speech signal that is pre-emphasized and hamming-windowed, its power spectrum can be formulated as

$$Y_i(k)=\left|\mathcal{F}[y_i(n)]\right|^2=\left|\sum_{n=1}^{N} y_i(n)\exp(-j\frac{2\pi nk}{N})\right|^2 \qquad (2)$$

where $1 \le k \le K$, K is the length of FFT.

If noise and speech signal are uncorrelated, the above equation can be further expressed as

$$Y_i(k)=X_i(k)+N_i(k) \qquad (3)$$

In the derivation of conventional mel-scaled frequency cepstral coefficients, the power spectrum is transformed into some coefficients in cepstral domain by passing it through a mel-scaled frequency filter bank whose outputs followed by a log operation and a DCT.

In this paper, we introduce another representation called differential power spectrum (DPS) which is defined by following differential equation:

$$D_i(k)=\sum_{l=-O}^{P} b_l Y_i(k+l) \qquad (4)$$

where $D_i(k)$ is the differential power spectrum of the ith frame of speech signal. P and O are the orders of the differential equation, and b_l's weighting coefficients.

Fig. 1 plots a frame of speech signal taken from the TI46 database, the power spectrum of this frame signal and its differential power spectrum. We see from this figure that if P, O and b_l's are properly selected, the spectral peaks are retained in the DPS, except that each peak being split into two, one positive and one negative. The flat part of power spectrum however, is transformed into some values approximating to zero. This observation interests us to investigate the DPS for speech recognition since spectral peaks convey the most important information in speech signal. The DPS preserves spectral peaks means that it does not lose much information contained in speech. On the other hand, noise spectrum is often quite flat. The differentiation operation will make the flat part of the spectrum to be near zero. Hence we can expect that DPS based representation is robust with respect to the noise whose spectrum is flat. In what follows, we will investigate the use of DPS for speech recognition.

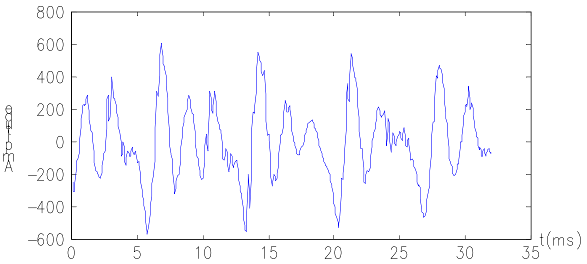

Fig. 1 (a) One frame of speech signal

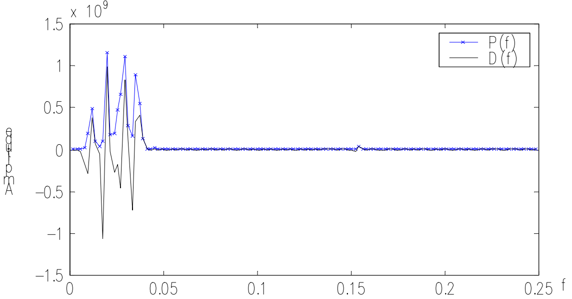

Fig .1 (b) The power spectrum and DPS of signal in (a)

The DPS is estimated via $D_i(k)=Y_i(k)-Y_i(k+1)$

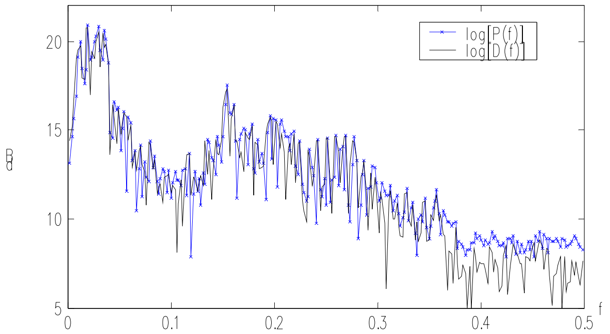

Fig .2 The power spectrum and magnitude square of the DPS of signal in (a)

The DPS is estimated via $D_i(k)=Y_i(k)-Y_i(k+1)$

2.2 Representing the DPS into Speech Features

Before the use of the DPS for speech recognition, we have to solve three problems. The first problem is the selection of proper orders of the differential equations, namely the P and O parameters in Eq. (4). The second one is the determination of weights b_l's in Eq. (4). The third one is how to convert the DPS into a few parameters that can best reflect information contained in speech signal, which is necessary for recognition purpose.

Unfortunately, an optimal solution to any of aforementioned three problems is difficult to achieve. Rather than seeking some criterions to optimize these problems, we will show only empirical solutions for practical applications.

For the first two problems, we will investigate and compare the use of following three special forms of DPS:

DPS1: $D_i(k)=Y_i(k)-Y_i(k+1)$ (5)

DPS2: $D_i(k)=Y_i(k)-Y_i(k+2)$ (6)

DPS3: $D_i(k)=Y_i(k-2)+Y_i(k-1)-Y_i(k+1)-Y_i(k+2)$ (7)

The third problem is circumvented by converting the DPS into some MFCC-like parameters. First, an absolute operation is applied to the DPS to make its negative parts positive. Fig. 2 plots the magnitude and the power spectrum of the signal presented in Fig. 1 (a). One can see that the magnitude of DPS has quite similar envelope with the power spectrum. This may indicate that the magnitude of the DPS preserves spectral shape information. Second, the magnitude of the DPS is passed through a mel-spaced frequency filter bank whose outputs are followed by a log operation. Finally, the logarithmic filter bank outputs are compressed into a feature vector with much lower dimensionality using a DCT. We refer this feature vector as DPS-based MFCC. For simplicity, we denote them as DPS without introducing any confusion.

3. Experiments

Various experiments have been performed to assess the proposed features and to compare them with the conventional MFCCs. For brevity, we cite some of them in this paper.

3.1 Isolated Speech Recognition

The first experiment uses the TI46 database to find which forms of DPS can lead to a better recognition performance. The TI46 is an isolated spoken words database which is designed and collected by Texas Instruments (TI). The database contains 16 speakers including 8 males and 8 females. The vocabulary consists of 10 isolated digits from ‘ZERO’ to ‘NINE’, 26 isolated English alphabets from ‘A’ to ‘Z’, and ten isolated words including “ENTER, ERASE, GO, HELP, NO, RUBOUT, REPEAT, STOP, START, YES”. There are 26 utterances of each word from each speaker: 10 of them are designated as training and the rest 16 are designated as testing tokens. Speech signal is digitized at a sampling rate of 12.5kHz with 12-bit quantization value for each sample.

In this experiment, we take speech from 8 male speakers to perform English alphabet recognition. Four sets of features are considered, namely MFCC, DPS1, DPS2, and DPS3.

MFCC: Speech signal is analyzed every 10ms with a frame width of 32ms (with preemphasis and Hamming windowing). For each frame, the FFT is performed to estimate its power spectrum, which is then fitted to a mel-scaled filter bank which consists of 24 triangular filters. 12 MFCCs are computed by applying a log operation and a cosine transform to the 24 filter bank energies (The first order cepstral coefficient C_0 is ignored).

DPS1: Speech signal is split into frames same as above. For each frame, the power spectrum is estimated using FFT. The differential power spectrum is then calculated according to Eq. 5. The magnitude of the DPS is then input to a same mel-scaled filter bank and converted it to 12 cepstral coefficients. Similarly, we compute the **DPS2** and **DPS3** according to Eq. 6 and Eq. 7 respectively.

The recognition system used is a speaker-independent whole-word-model based HMM recognizer. Models are left-to-right with no skip state transition. Eight states are used for each model. A mixture of 4 multivariate Gaussian distributions with diagonal covariance matrices is used for each state to approximate its probability density function. The training iterations begin with a uniform segmentation. Experiment results are shown in Table 1.

Table 1: Word accuracy using different feature sets

	MFCC	DPS1	DPS2	DPS3
12 S	84.05	86.11	85.11	85.69
12 S + 12 D	90.39	92.11	90.87	92.08
12 S + 12 D + 12 A	91.75	93.63	91.72	92.78

(S: static features; D: dynamic features; A: accelerations)

From above results, we can see that all three kinds of DPS based feature vectors are superior to the conventional MFCCs. Among the three DPS definitions, the DPS1 defined by Eq. 5 yields the best performance.

In the subsequent experiments, we will evaluate the DPS based features and its robustness with respect to noise for various recognition tasks. As we have shown that the DPS1 can yield more promising results, we will only assess the DPS1 based features. We shall compare them with the conventional MFCCs.

3.2 Connected digits recognition

The second experiment is to recognize connected digits. The TI connected digits database [16] is used for this purpose. This database contains digit strings uttered by adult speakers and children as well. However, only digit strings from 225 adult talkers are used in this experiment. These strings are originally divided into training set and test set for consistency of comparison of results among different researchers.

The vocabulary in this database consists of 11 words which include 10 digits and an “oh”. Each talker uttered 77 sequences of these words, consisting of 2 tokens of each of the 11 words in isolation, and 11 strings of each of 2, 3, 4, 5, and 7 digits. The digit strings were recorded in an acoustically treated sound room with a sampling frequency being 20 kHz. For the comparison with the recognition results reported, we downsampled speech to 8 kHz using Matlab downsampling function.

To test the robustness of different front-ends with respect to noise, we directly add some noise to the speech signal in the test set. The training speech is kept clean. The noise signals used are from NOISEX database [17]. The noise signal provided in this database is sampled at 16kHz. To match its bandwidth to the speech signal, we downsampled the noise signal to 8 kHz.

The HTK speech recognition system is used to perform the recognition task. This was configured as a gender-independent mixture Gaussian HMM system. The model set consists of 11 word-models, a silence model and a short pause model. Except the short pause, each model has 6 emitting states. The short pause model has only one emitting state. A mixture of 8 multivariate Gaussian distributions with diagonal covariance matrices is used for each emitting state to approximate its probability density function.

Four sets of feature vectors are investigated in this experiment:

MFCC: Speech signal is analyzed every 15ms with a frame width of 32ms (with preemphasis and Hamming windowing). Each frame is transformed into 12 MFCCs using the procedure same as that in Experiment 1. Moreover, the normalized log frame energy is also added to the 12MFCCs to form a 13-dimentional static vector. This static vector is then expanded to produce a 39-dimentional feature vector (static + differentiation + acceleration).

DPS: Speech signal is split into frames as above. For each frame, the power spectrum is estimated and the differential power spectrum is then calculated according to Eq. 5. The magnitude of the DPS is converted to 12 MFCC-like coefficients. This 12-dimentional vector is further expanded to a 39-dimentional feature vector using same strategy as used to computed MFCC features.

MFCC*: MFCC + CMN (Cepstral Mean Normalization).

DPS*: DPS+CMN.

The experiment results are shown in Table 2.

Table 2 a: Word accuracy in speech noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	18.8	41.3	72.6	91.6	97.2	98.7	99.0
DPS	22.6	44.2	76.7	92.4	97.60	98.6	99.0
MFCC*	20.7	46.7	77.2	92.7	96.8	98.1	98.8
DPS*	29.9	60.3	85.3	94.7	97.2	98.5	98.9

Table 2 b: Word accuracy in machine-gun noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	80.2	87.6	93.6	97.3	98.4	98.9	99.0
DPS	84.31	91.38	96.14	98.06	98.7	99.0	99.0
MFCC*	81.9	88.8	94.2	97.4	98.5	98.8	98.8
DPS*	88.4	94.3	97.3	98.5	98.9	99.0	98.9

Table 2 c: Word accuracy in Lynx noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	28.6	56.8	82.1	93.9	97.5	98.7	99.0
DPS	27.7	55.87	79.9	94.3	97.9	98.6	99.0
MFCC*	26.9	57.0	83.9	94.6	97.1	98.2	98.8
DSP*	33.5	65.4	87.5	95.1	97.4	98.5	98.9

From Table 2, we can make following observations:

1. As compared with the conventional MFCCs, the new cepstral vector derived from the DPS yields at least comparable performance in clean as well as high SNR conditions.
2. In most strong noise conditions, the DPS feature set outperforms the MFCCs.
3. For machine-gun noise condition, the improvement of recognition performance is more significant. The reason for this is under investigation.
4. CMN is effective to augment the robustness of both MFCCs and DPS features with respect to noise.
5. After CMN, the DPS features outperform the conventional MFCCs in both clean and noisy conditions.

3.3 Phone Recognition

The second experiment is to perform phone recognition. The speech data employed in this experiment is the TIMIT phoneme based continuous speech database [20], which contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. This database is split into a training set of 3696 utterances and a test set which contains 1344 utterances. Speech signal is sampled at 16kHz with 16 bits per-word.

The TIMIT database is phonetically transcribed using a set of 61 phones. To facilitate comparison with the results reported, we perform phonetic recognition on this database over the set of 39 classes that are commonly used for such evaluation [21]. Again, the HTK toolkit is configured to perform the recognition task. The model set consists of 39 mono-phone HMMs. Each model has three emitting states. An 8-component mixture Gaussian distribution is used for each emitting state to approximate the probability density function. Phoneme bigram is used as a language model.

We assess two feature sets:

MFCC*: MFCC + CMN (39 coefficients).

DPS*: DPS+CMN (39 coefficients).

The static MFCCs and DPS based cepstral coefficients are estimated using the same procedure as described in the previous experiment. Only difference is that analysis frame length in this experiment is 32 ms with 10 ms overlap.

Recognition results for this experiment are shown in Table 3.

Table 3 a: Phone accuracy in speech noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC*	22.36	41.99	55.75	61.42	62.72	63.01	63.00
DPS*	23.65	43.86	56.41	61.56	62.57	62.87	62.97

Table 3 b: Phone accuracy in machine-gun noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC*	55.13	59.32	61.37	62.50	62.86	63.00	63.00
DPS*	55.63	59.80	61.76	62.40	62.77	62.89	62.97

Table 3 c: Word accuracy in Lynx noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC*	29.10	47.31	58.27	62.12	62.82	63.03	63.00
DPS*	30.15	48.85	58.68	61.74	62.60	62.97	62.97

From the above table, we see that:

1. Compared with the conventional MFCCs, the DSP features yield comparable results in clean and weak noise conditions.
2. The DPS based cepstrum slightly outperforms the conventional MFCC vector in strong noise conditions.

3.4 Evaluation of AURORA Task

The AURORA task [22] has been defined by the European Telecommunications Standards (ETSI) as a cellular industry imitative to standardize a robust feature extraction technique for a distributed speech recognition framework. This task used the TIDigits database downsampled from the original sampling frequency of 20kHz to 8 kHz with an “ideal” low-pass filter and normalized to the same amplitude level.

To account for the realistic frequency characteristics of terminals and equipment in the telecommunication area, an additional filtering is applied. The two “standard” frequency characteristics used are G.712 and MIRS [22].

Noise is artificially added to the filtered TIDigits at SNRs of 20dB, 15dB, 10dB, 5dB, 0dB and –5dB. Noise signals are recorded at different places including suburban train, crowed of people (babble), car, exhibition hall, restaurant, street, airport and training station.

Two training modes are defined, i.e., training on clean data only and training on clean as well as noisy data (multi-condition). For the first mode, 8440 utterances are selected from the training part of the TIDigits containing the recording of 55 male and 55 female adults. These signals are filtered with the G. 712 characteristic without noise added. For the second mode, 8440 utterances from TIDigits training part are equally split into 20 subsets with 422 utterances in each subset. Each subset contains a few utterances of all training speakers. Suburban train, babble, car, and exhibition hall noises are added to 20 subsets at 5 different SNRs, namely, 20dB, 15dB, 10dB, 5dB and the clean condition. Both speech and noise are filtered before adding.

Three test sets are defined. 4004 utterances from 52 male and 52 female speakers in the TIDigits test part are divided into 4 subsets with 1001 utterances in each. Recordings from all speakers are present in each subset. One noise is added to each subset at SNRs of 20dB to –5dB in decreasing steps of 5dB after speech and noise being filtered with the G. 712. The three subset are as below:

Test Set A: Suburban train, babble, car and exhibition noise are added to the 4 subsets. In total, this set contains $4 \times 7 \times 1001$ utterances. This set leads to a high match of training and test data as it contains same noises as used for the multi-condition trainings mode.

Test Set B: It is created in exactly same way, but using the four different noises, namely, restaurant, street, airport and train station.

Test Set C: It contains 2 of the 4 subsets. Speech and noise are filtered with the MIRS characteristic before adding. Two types of noise, i.e., suburban train and street noise, are added at 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB.

To facilitate comparison of results among different researchers, Aurora task provides a reference recognizer which is based on the HTK software package. The models set contains 11 whole word HMMs and two pause models, i.e., “sil” and “sp”. Each word model has 16 states with each state having 3 mixtures. “sil” model has 3 states and each state has 6 mixtures. “sp” has only single state.

Aurora task also provide a baseline performance which uses conventional MFCCs and MFCCs after CMN as front-end features. The details for calculating MFCCs are shown as below:

1. Frame length is 25ms. Frame shift is 10ms.
2. Preemphasis with a factor of 0.97.
3. Application with a Hamming window.
4. FFT based Mel filter bank with 23 frequency bands in the range from 64 Hz up to 4kHz

The logarithmic frame energy is added to the 12MFCCs (the MFCC of order 0 is ignored) to construct 13-dimentional static feature vector. This vector is further expanded to a 39-dimensional vector by including its delta and acceleration coefficients.

We assess the DPS features on this task. The DPS based cepstral coefficients are computed in exactly the same way except that our Mel filter bank consists of 24 frequency bands rather than 23 bands. We also include the logarithmic frame energy to augment recognition performance. The final feature vector also contains 39 coefficients including 13 static, 13 delta and 13 acceleration coefficients.

The word recognition accuracies for three test sets in different noisy conditions using different feature vectors are shown in Table 4. The average accuracies and the error rate reduction as compared with baseline MFCCs system are also provided. From the results, we can make the following observations:

1. Comparing test set A and B in Table 4 (b) with those in 4 (a), one can find that the CMN can improve the robustness of MFCCs with respect to additive noise, as no channel distortion is introduced in these two sets.
2. With the use of CMN, the error rate is reduced 4.7%, 7.7% and 16.8% for test set A, B and C respectively. That the error reduction for set C is much more significant shows the effectiveness of the CMN in dealing with channel effects since only set C is corrupted both by channel distortion and by additive noise.
3. Using the DPS based cepstral coefficients and CMN, we get an overall error reduction of 21.7% for the whole task and of 15.8% for set A, 17.9% for set B and 36.8% for test C. This justifies the effectiveness of the new feature set.
4. Again, the performance improvement for set C is more significant than those for set A and B. We owe this achievement to the CMN technique.

Table 4 a: Word accuracy for baseline system using MFCCs and logarithm frame energy as front-ends (39-dimentionsal feature vector)

	A					B					C			
	Subway	Babble	Car	Exhib.	Average	Rest.	Street	Airport	Station	Average	Sub.	Street	Average	**Overall**
Clean	98.68	98.52	98.39	98.49	**98.52**	98.68	98.52	98.39	98.49	**98.52**	98.50	98.58	**98.54**	**98.52**
20 dB	97.61	97.73	98.03	97.41	**97.70**	96.87	97.58	97.44	97.01	**97.23**	97.30	96.55	**96.93**	**97.35**
15 dB	96.47	97.04	97.61	96.67	**96.95**	95.30	96.31	96.12	95.53	**95.82**	96.35	95.53	**95.94**	**96.29**
10 dB	94.44	95.28	95.74	94.11	**94.89**	91.96	94.35	93.29	92.87	**93.12**	93.34	92.50	**92.92**	**93.79**
5 dB	88.36	87.55	87.80	87.60	**87.83**	83.54	85.61	86.25	83.52	**84.73**	82.41	82.53	**82.47**	**85.52**
0 dB	66.90	62.15	53.44	64.36	**61.71**	59.29	61.34	65.11	56.12	**60.47**	46.82	54.44	**50.63**	**59.00**
-5dB	26.13	27.18	20.58	24.34	**24.56**	25.51	27.60	29.41	21.07	**25.90**	18.91	24.24	**21.58**	**24.50**
Average	**88.76**	**87.95**	**86.52**	**88.03**	**87.82**	**85.39**	**98.52**	**87.64**	**98.49**	**86.27**	**83.24**	**84.31**	**83.78**	**86.39**

Table 4 b: Performance for MFCCs + CMN (39-dimentionsal feature vector)
(ERR: Error rate reduction in comparison with baseline performance)

	A					B					C				
	Subway	Babble	Car	Exhib.	Average	Rest.	Street	Airport	Station	Average	Sub.	Street	Average	**Overall**	**ERR**
Clean	98.68	98.91	98.60	98.55	**98.69**	98.68	98.91	98.60	98.55	**98.69**	98.89	98.88	98.89	**98.73**	**13.39%**
20 dB	97.85	97.88	98.00	97.13	**97.72**	97.33	97.73	97.73	97.59	**97.60**	97.91	97.40	97.66	**97.66**	**10.32%**
15 dB	96.93	97.31	97.44	96.45	**97.03**	95.70	96.67	96.36	95.74	**96.12**	96.99	96.70	96.85	**96.63**	**8.13%**
10 dB	95.21	95.44	95.91	93.00	**94.89**	92.20	95.07	94.39	93.74	**93.85**	94.26	93.92	94.09	**94.31**	**7.95%**
5 dB	89.90	88.24	88.40	86.33	**88.22**	85.29	87.27	88.25	85.22	**86.51**	86.21	83.86	85.04	**86.90**	**8.97%**
0 dB	70.40	63.21	56.16	66.58	**64.09**	61.50	62.88	67.97	57.85	**62.55**	61.07	56.62	58.85	**62.42**	**7.86%**
-5dB	32.70	27.96	20.49	28.82	**27.49**	28.55	27.72	31.40	21.69	**27.34**	28.28	26.36	27.32	**27.40**	**3.80%**
Average	**90.06**	**88.42**	**87.18**	**87.90**	**88.39**	**86.40**	**87.92**	**88.94**	**86.03**	**87.32**	**87.29**	**85.70**	**86.49**	**87.58**	
ERR	**11.58%**	**3.87%**	**4.88%**	**-1.10%**	**4.71%**	**6.93%**	**6.84%**	**10.50%**	**6.79%**	**7.67%**	**24.13%**	**8.86%**	**16.75%**		**8.77%**

Table 4 c: Performance for DPS cepstral features + CMN (39-dimentionsal feature vector)
(ERR: Error rate reduction compared with baseline performance)

	A					B					C				
	Subway	Babble	Car	Exhib.	Average	Rest.	Street	Airport	Station	Average	Sub.	Street	Average	**Overall**	**ERR**
Clean	99.08	98.76	99.11	99.26	**99.05**	99.08	98.76	99.11	99.26	**99.05**	99.05	98.85	98.95	**99.03**	**34.01%**
20 dB	97.94	98.31	98.54	98.03	**98.21**	98.04	98.00	98.42	98.67	**98.28**	98.13	97.64	97.89	**98.17**	**30.01%**
15 dB	97.39	97.70	97.88	96.98	**97.49**	97.36	97.22	97.55	97.13	**97.32**	97.64	96.74	97.19	**97.36**	**27.25%**
10 dB	95.70	96.34	95.68	94.72	**95.61**	94.57	95.19	95.38	93.98	**94.78**	95.33	94.32	94.83	**95.12**	**20.23%**
5 dB	91.46	88.91	90.46	87.44	**89.57**	85.42	88.60	89.11	86.02	**87.29**	90.42	88.00	89.21	**88.58**	**20.31%**
0 dB	75.13	64.93	62.48	68.74	**67.82**	62.67	68.20	71.25	61.83	**65.99**	72.09	67.20	69.65	**67.45**	**19.61%**
-5dB	34.05	27.45	20.46	29.10	**27.77**	27.48	28.93	31.35	21.85	**27.40**	32.42	27.93	30.18	**28.10**	**4.70%**
Average	**91.52**	**89.24**	**89.01**	**89.18**	**89.74**	**87.61**	**89.44**	**90.34**	**87.53**	**88.73**	**90.72**	**88.78**	**89.75**	**89.34**	
ERR	**24.62%**	**10.69%**	**18.43%**	**9.62%**	**15.78%**	**15.20%**	**18.55%**	**21.85%**	**16.78%**	**17.92%**	**44.63%**	**28.49%**	**36.82%**		**21.66%**

4. CONCLUSION

In this paper, a new set of cepstral feature vector derived from the differential power spectrum is introduced. Experiments for isolated, connected and continuous speech recognition tasks show that the new MFCCs yield at least comparable performance as the conventional MFCCs in both clean and noise conditions. In most cases, this new feature set outperforms the conventional MFCCs.

References

[1] S. Nicholson, *et al*, “Evaluating feature set performance using the F-ratio and J-measures,” in Proc. EUROSPEECH’97, PP. 413-416.

[2] K. K. Paliwal, “Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer,” Digital signal processing, Vol. 2, PP.157-173. 1992.

[3] S. V. Vaseghi, *et al*, “Noise Compensation Methods for Hidden Markov Model Speech Recognition in Adverse Environments,” IEEE Trans. Speech and Audio Processing, Vol. 5, No. 1, Jan. 1997. PP. 11-21.

[4] D. C. Popescu, *et al*, “Kalman Filtering of Colored Noise for Speech Enhancement,” in Proc. ICASSP’98, PP. 997-1000.

[5] S. F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 27, No. 2, April 1979. PP. 113-120.

[6] H. G. Hirsch, *et al*, “Improved speech recognition using high-pass filtering of subband envelopes,” in Proc. EUROSPEECH, PP. 413-416, 1991.

[7] D. Geller, *et al* ,“Improvements in speech recognition for voice dialing in the car environment,” in Proc. ESCA Workshop on Speech Processing in Adverse Conditions, PP. 203-206, Nov. 1992.

[8] M. Rahim, and B. -H. Juang, “Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition”, IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, PP. 19-30, January 1996.

[9] J. C. Junqua *et al*, “Environment-adaptive algorithms for robust speech recognition,” in Proc. HSC2001, PP.31-34, April 9-11, Kyoto, Japan

[10] M. J. F. Gales, *et al*,, “Robust speech recognition using parallel model combination,” IEEE Trans. Speech Audio Processing, Vol. 4, PP. 352-359, Sep. 1996.

[11] P. J. Moreno, *et al*, “A vector Taylor series approach for environment independent speech recognition,” in Proc. ICASSP’96, May 1996, PP.733-736.

[12] S. Sagayama et al, “Jacobian Approach to Fast Acoustic Model Adaptation,” in Proc. ICASSP’97, PP. 835-838.

[13] P. C. Woodland, *et al*, “Improving environmental robustness in large vocabulary speech recognition,” in Proc. ICASSP’96, May 1996, PP. 65-68.

[14] K. Ohkura, *et al*, “Speaker Adaptation Based on Transfer Vector Filed Smoothing Technique,” in Proc. ICSLP’1992, PP. 369-372.

[15] J. A. Nolazco Flores, *et al*, “Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation”, in Proc. ICASSP’94, Vol. I, PP. 409-412.

[16] http://www.ldc.upenn.edu/readme.files/tidigits.readme.html

[17] A. Varga, *et al*, “The Noise-92 Study on the Effect of Additive Noise on Automatic Speech Recognition,” DRA Speech Research Unit, St. Andrew’s Rd., Malvern, Worcestershire, WR14 3PS UK.

[18] H. Bourlard, *et al*, “A mew ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands,” in Proc. ICSLP’96, Philadelphia, October 1996.

[19] J. R. Cohen, “Application of an Auditory Model to Speech Recognition,” J. Acoustic. Soc. Amer., Vol. 85, June 1989, PP. 2623-2329.

[20] L. F. Lamel, *et al*, “Speech database development: Design and analysis of the acoustic-phonetic corpus,” in Proc. DARPA Speech Recognition Workshop, PP. 100-109, Feb. 1986.

[21] K. F. Lee, et al, “Speaker-Independent Phone Recognition Using Hidden Markov Models”, IEEE Trans. ASSP, Vol. 37, No. 11, Nov. 1989. PP. 1641-1648.

[22] H. G. Hirsch, “The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under noisy Conditions”, In Proc. ISCA ASR2000, Paris, France, Sep. 2000.