Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Speech Communication 41 (2003) 469–484

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Cepstrum derived from differentiated power spectrum for robust speech recognition ☆

Jingdong Chen [a,*], Kuldip K. Paliwal [b], Satoshi Nakamura [c]

[a] *Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974-0636, USA*
[b] *School of Microelectronic Engineering, Griffith University, Brisbane QLD 4111, Australia*
[c] *ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288, Japan*

Received 15 January 2002; accepted 5 January 2003

Abstract

In this paper, cepstral features derived from the differential power spectrum (DPS) are proposed for improving the robustness of a speech recognizer in presence of background noise. These robust features are computed from the speech signal of a given frame through the following four steps. First, the short-time power spectrum of speech signal is computed from the speech signal through the fast Fourier transform algorithm. Second, DPS is obtained by differentiating the power spectrum with respect to frequency. Third, the magnitude of DPS is projected from linear frequency to the mel scale and smoothed by a filter bank. Finally, the outputs of the filter bank are transformed to cepstral coefficients by the discrete cosine transform after a nonlinear transformation. It is shown that this new feature set can be decomposed as the superposition of the standard cepstrum and its nonlinearly liftered counterpart. While a linear lifter has no effect on the continuous density hidden Markov model based speech recognition, we show that the proposed feature set embedded with a nonlinear liftering transformation is quite effective for robust speech recognition. For this, we conduct a number of speech recognition experiments (including isolated word recognition, connected digits recognition, and large vocabulary continuous speech recognition) in various operating environments and compare the DPS features with the standard mel-frequency cepstral coefficient features used with cepstral mean normalization and spectral subtraction techniques.
© 2003 Elsevier B.V. All rights reserved.

Keywords: Robust speech recognition; Hidden Markov model; Differential power spectrum; Linear liftering; Cepstral mean normalization; Spectral subtraction

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 crucial step is the feature extraction, where the speech signal of a given frame is converted to a set of acoustic features with the hope that these features will encapsulate the important information that is necessary for recognition. Once these features are computed, a back-end classifier is used to recognize the input speech signal into a sequence of words in light of the extracted features and pre-trained models.

☆ This work was carried out while the first author was in ATR Spoken Language Translation Research Laboratories.

* Corresponding author. Tel.: +1-908-582-5044; fax: +1-908-582-7308.
E-mail address: jingdong@research.bell-labs.com (J. Chen).

0167-6393/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-6393(03)00016-5

Acoustic features may greatly affect the performance of a speech recognizer. Three criteria can be used to evaluate a feature set. These are: discriminability, robustness and complexity. The discriminability requires the selected feature set to have capability to discriminate among different acoustic units. Though this can be evaluated through separability, a more pertinent way is to assess through speech recognition in matched conditions. The robustness demands the feature set to be resilient to acoustical distortions such as additive noise and convolutive channel effect. This can be tested through recognition in degraded conditions. Since an ASR system is expected to perform real time recognition, a computationally efficient algorithm to extract the features is essential. This is often examined by computational complexity of the feature extraction process.

A great deal of work has been done for feature extraction (Davis and Mermelstein, 1980; Furui, 1986; Soong and Rosenberg, 1986; Picone, 1993; Hermansky, 1990; Bourlard and Dupont, 1996; Kim et al., 1999). Among those features that have been investigated and reported, the mel-frequency cepstral coefficients (MFCCs), developed by Davis and Mermelstein (1980), are used almost as "standard" acoustic parameters in currently available speech recognition systems. Much evidence shows that MFCCs have served as very successful front-ends for the hidden Markov model (HMM) based speech recognition in the past decade. Many speech recognition systems based on these front-ends have achieved a very high level of accuracy in clean speech environment.

Despite the de facto standardization of their use as front-ends, MFCCs are widely acknowledged not to cope well with noisy speech. In the literature, various approaches have been proposed to improve the tolerance of an ASR system with respect to noise, such as Wiener filtering (Vaseghi and Milner, 1997), Kalman filtering (Popescu and Zeljkovic, 1998), spectral subtraction (Boll, 1979; Nolazco Flores and Young, 1994), RASTA (Hermansky et al., 1991; Hirsch et al., 1991), lin-log RASTA (Hermansky and Morgan, 1994), cepstral mean removal (Geller et al., 1992), signal bias removal (Rahim and Juang, 1996), parallel model compensation (PMC) (Gales and Young, 1996), vector Taylor series approximation based model compensation (Moreno et al., 1996), Jacobian approach (Sagayama et al., 1997; Junqua et al., 2001), maximum likelihood linear regression (MLLR) (Woodland et al., 1996), and transfer vector interpolation (Ohkura et al., 1992), to name a few. These methods often take advantage of the prior knowledge of noise to mask, cancel or remove noise during front-end processing or adjust the system parameters to match the new noisy environment to improve recognition performance.

Although the aforementioned efforts were experimented in speech recognition with certain success, there remains a great need to investigate new technologies to improve the basic ASR in order to meet the high performance objectives set for practical speech recognition applications. To improve the performance of modern ASR systems, it is crucial to develop new features set since all the succeeding processing in ASR systems are highly dependent on the quality of the extracted features.

In this paper, we present a new set of cepstral coefficients derived from the differential power spectrum (DPS) for speech recognition. First, the short-time power spectrum of speech signal is estimated through FFT. The power spectrum estimate is then differentiated with respected to frequency. Finally, the magnitude of DPS is converted to some coefficients in the cepstral domain by passing it through a mel-frequency filter bank whose outputs are followed by a nonlinear transformation and DCT.

We show that the new cepstrum can be expressed as the superposition of the conventional cepstrum and its nonlinearly liftered counterpart. While a linear liftering transform has no effect on continuous density HMM-based speech recognition, experiments for various recognition tasks in different noise conditions indicate that the proposed feature set, which is embedded with a nonlinear lifter, is more tolerant to noise when compared to MFCCs. Experiment is also performed to compare the new feature with the widely used spectral subtraction technique.

2. Cepstrum derived from the differential power spectrum

2.1. Definition of the differential power spectrum

If denoted by $s(t)$ the original clean speech signal, the received speech signal $y(t)$ can be modeled as

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

where $h(t)$ represents the impulse response of the transmission channel, $*$ indicates the convolution operator, $v(t)$ is the ambient noise, and $x(t)$ the noise-free speech signal.

Speech signal is time-variant and nonstationary. It is usually analyzed on the frame-by-frame basis. If we assume that $y(n) = x(n) + v(n)$ $(0 \leqslant n < N$, where N is the frame length) represents a given frame of a speech signal that is pre-emphasized and hamming-windowed, its power spectrum can be formulated as

$$Y(\omega) = \mathscr{F}[r_y(\tau)] = \sum_{\tau=-N+1}^{N-1} r_y(\tau)\mathrm{e}^{-\mathrm{j}\omega\tau}, \tag{2}$$

where $\mathscr{F}[\cdot]$ indicates the Fourier transform, ω denotes radian frequency, and $r_y(\tau)$ is the short-time autocorrelation sequence which is given as

$$r_y(\tau) = \begin{cases} \frac{1}{N}\sum_{k=0}^{N-\tau-1} y(k)y(k+\tau), & \text{for } \tau = 0, 1, \ldots, N-1, \\ r_y(-\tau), & \text{for } \tau = -N+1, -N+2, \ldots, -1. \end{cases} \tag{3}$$

If we assume that the noise and speech signal are mutually uncorrelated, (2) can be recast as

$$\begin{aligned} Y(\omega) &= \mathscr{F}[r_y(\tau)] \approx \mathscr{F}[r_x(\tau)] + \mathscr{F}[r_v(\tau)] \\ &= X(\omega) + V(\omega). \end{aligned} \tag{4}$$

In current speech recognition systems, the power spectrum is often represented into some cepstral coefficients through the following transformation:

$$\begin{aligned} c(m) &= \mathscr{F}^{-1}[\log Y(\omega)] \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} [\log Y(\omega)]\mathrm{e}^{\mathrm{j}\omega m}\,\mathrm{d}\omega. \end{aligned} \tag{5}$$

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

$$D(\omega) = Y'(\omega) = \frac{\mathrm{d}Y(\omega)}{\mathrm{d}\omega}, \tag{6}$$

where the prime represents differentiation with respect to ω. If the uncorrelation assumption holds, we see that, using (4),

$$\begin{aligned} D(\omega) &= \frac{\mathrm{d}Y(\omega)}{\mathrm{d}\omega} = \frac{\mathrm{d}X(\omega)}{\mathrm{d}\omega} + \frac{\mathrm{d}V(\omega)}{\mathrm{d}\omega} \\ &= D_X(\omega) + D_V(\omega), \end{aligned} \tag{7}$$

where $D_X(\omega)$ and $D_V(\omega)$ are the differential power spectra of the given frame of noise-free speech and noise signal, respectively. This definition of DPS is given in the continuous frequency domain. Its discrete counterpart can be approximated in terms of following difference equation:

$$\begin{aligned} D(k) &\approx \sum_{l=-O}^{P} b_l Y(k+l) \\ &\approx \sum_{l=-O}^{P} b_l [X(k+l) + V(k+l)] \\ &= D_X(k) + D_V(k), \end{aligned} \tag{8}$$

where P and O are the orders of the differential equation, b_l's some real-valued weighting coefficients, and $0 \leqslant k < K$, here K is the length of FFT.

Fig. 1 plots a frame of speech signal taken from the TI46 database (see Section 3), its power spectrum and DPS. It can be seen from this figure that for the selected difference equation, the spectral peaks are retained in the DPS representation, except that each peak is split into two, one positive and one negative. The flat part of the power spectrum however, is transformed into some values approximating to zero. This interesting observation motivates us to investigate DPS for speech recognition since spectral peaks convey the most important information in speech signal. The fact that DPS preserves spectral peaks means that the DPS representation does not lose information contained in the speech signal. On the other hand, noise spectrum is often quite flat. The differentiation operation will cause 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,

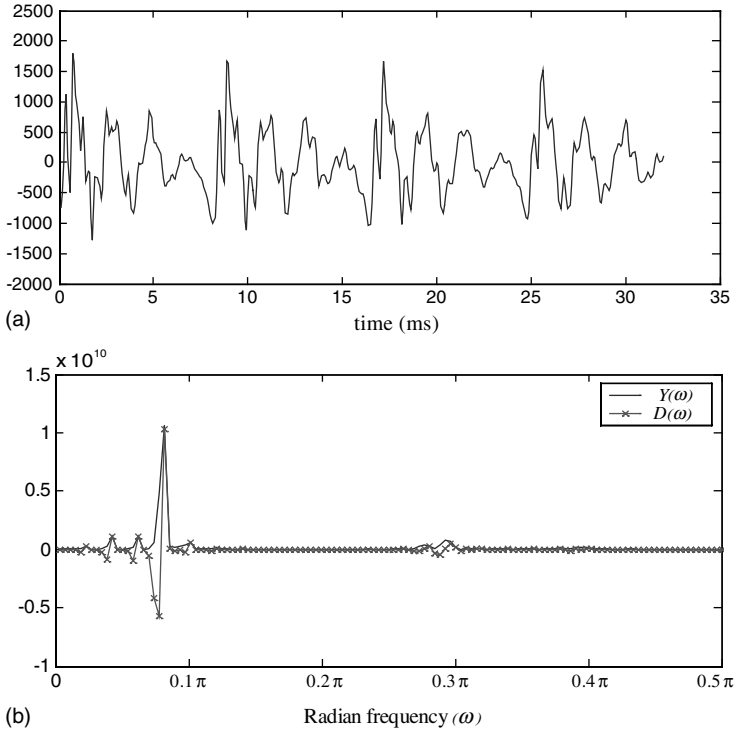

Fig. 1. The power spectrum and DPS for an 'a' sound from the TI46 database: (a) waveform plot of a frame of the 'a' sound; (b) the power spectrum and DPS of the signal shown in (a). (The power spectrum is estimated through 512-point FFT. DPS is computed via $D(k) = Y(k) - Y(k+1)$. The radian frequency varies from 0 to π. Only range between 0 and 0.5π is shown here to enable a clear view of peaks.)

we will investigate the use of DPS for speech recognition.

2.2. Representing DPS into speech features

Before the use of DPS for speech recognition, we have to resolve three problems. The first one is the selection of proper orders of the difference equations, namely the P and O parameters in (8). The second one is the determination of weights b_l's in (8). The third one is how DPS should be converted into a few parameters that can best reflect information contained in a speech signal, which is necessary for recognition purpose.

Unfortunately, an optimal solution to any of the three listed problems is difficult to achieve. Rather than seeking some criteria 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:

$$\text{DPS1:} \quad D(k) = Y(k) - Y(k+1), \tag{9}$$

$$\text{DPS2:} \quad D(k) = Y(k) - Y(k+2), \tag{10}$$

$$\text{DPS3:} \quad D(k) = Y(k-2) + Y(k-1) - Y(k+1) - Y(k+2). \tag{11}$$

The third problem is circumvented by converting DPS into cepstral coefficients. First, an absolute operation is applied to DPS to make its negative parts positive. Fig. 2 shows the magnitude of DPS and the power spectrum of the frame of speech signal presented in Fig. 1(a). One can see that the magnitude of DPS has an envelope quite similar to that of the power spectrum. This may

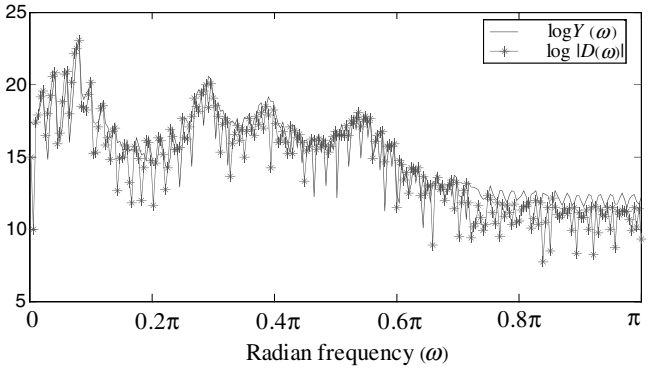

Fig. 2. The power spectrum and magnitude of DPS of the signal in Fig. 1(a). (The power spectrum and DPS are calculated in the same way as shown in Fig. 1.)

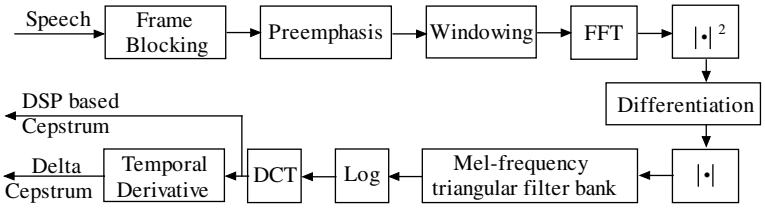

Fig. 3. Schematic diagram to extract DPSCCs.

indicate that the magnitude of DPS preserves spectral shape information. Second, the magnitude of DPS is passed through a mel-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 DCT. In summary, the schematic procedure to extract the new feature is shown in Fig. 3. We refer this feature as DPS-based cepstral coefficients. For simplicity, we denote them as DPSCC.

2.3. Comparison with the cepstral liftering technique

If the cepstral coefficients for a given frame of speech signal is denoted as $c(m)$, $m = 1, 2, \ldots, D$, here D is the dimension of the feature vector, then the corresponding liftered cepstral features are defined by

$$\varsigma(m) = w(m)c(m), \tag{12}$$

where $w(m)$, $m = 1, 2, \ldots, D$, defines the lifter. In a more compact matrix-vector form, (12) can be rewritten as

$$\vec{\varsigma} = W\vec{c}, \tag{13}$$

where $\vec{\varsigma} = [\varsigma(1), \varsigma(2), \ldots, \varsigma(D)]^{\mathrm{T}}$ and $\vec{c} = [c(1), c(2), \ldots, c(D)]^{\mathrm{T}}$ are the liftered and original cepstral vectors, respectively. Here, the symbol T denotes vector or matrix transpose, and

$$W = \begin{bmatrix} w(1) & 0 & \cdots & 0 \\ 0 & w(2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w(D) \end{bmatrix} \tag{14}$$

is the linear liftering transform.

The liftering technique was well investigated in the 70's and early 80's for speech recognition (Paliwal, 1992, 1999; Tohkura, 1987; Juang et al., 1987). It has been found to give significant recognition gains for dynamic time warping based speech recognition systems.

For continuous density HMM-based speech recognizer, however, the linear liftering transform has been shown to be ineffective (Paliwal, 1999). This is due to the fact that a linear transformation does not affect the logarithmic likelihood score. In fact, the likelihood of the vector $\vec{c}$ for a given class is computed through the following Mahalanobis distance

$$d(\vec{c},\vec{\mu}_c,\Sigma_c)=(\vec{c}-\vec{\mu}_c)^{\mathrm{T}}\Sigma_c^{-1}(\vec{c}-\vec{\mu}_c), \tag{15}$$

where $\vec{\mu}_c$ and Σ_c are the mean vector and covariance matrix, respectively, representing the given class, and obtained from training process.

When the liftered cepstral vector is used as a feature, the Mahalanobis distance is given by

$$d(\vec{\varsigma},\vec{\mu}_\varsigma,\Sigma_\varsigma)=(\vec{\varsigma}-\vec{\mu}_\varsigma)^{\mathrm{T}}\Sigma_\varsigma^{-1}(\vec{\varsigma}-\vec{\mu}_\varsigma). \tag{16}$$

Using (13), we can easily obtain

$$\begin{aligned}\vec{\mu}_\varsigma &= W\vec{\mu}_c,\\ \Sigma_\varsigma &= W\Sigma_c W^{\mathrm{T}}.\end{aligned} \tag{17}$$

Substituting (17) to (16) yields,

$$d(\vec{\varsigma},\vec{\mu}_\varsigma,\Sigma_\varsigma)=d(\vec{c},\vec{\mu}_c,\Sigma_c). \tag{18}$$

Therefore, the Mahalanobis distance, and eventually the logarithmic likelihood score, is invariant under the linear liftering transform. This proves that a linear lifter has no effect on continuous density HMM-based speech recognition.

Knowing that the cepstral coefficients derived from the power spectrum is expressed as

$$\begin{aligned}c(m)&=\mathscr{F}^{-1}[\log Y(\omega)]\\ &=\frac{1}{2\pi}\int_{-\pi}^{\pi}[\log Y(\omega)]\mathrm{e}^{\mathrm{j}\omega m}\,\mathrm{d}\omega,\end{aligned} \tag{19}$$

we can express from Fig. 3 the DPS based cepstrum as (neglect the filter bank analysis),

$$\begin{aligned}\eta(m)&=\mathscr{F}^{-1}[\log D(\omega)]\\ &=\frac{1}{2\pi}\int_{-\pi}^{\pi}[\log Y'(\omega)]\mathrm{e}^{\mathrm{j}\omega m}\,\mathrm{d}\omega.\end{aligned} \tag{20}$$

Using the fact that

$$\frac{\mathrm{d}\log Y(\omega)}{\mathrm{d}\omega}=\frac{Y'(\omega)}{Y(\omega)}, \tag{21}$$

we can rewrite (20) as

$$\begin{aligned}\eta(m)&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\{\log Y(\omega)+\log[\log Y(\omega)]'\}\mathrm{e}^{\mathrm{j}\omega m}\,\mathrm{d}\omega\\ &=\frac{1}{2\pi}\int_{-\pi}^{\pi}\log Y(\omega)\mathrm{e}^{\mathrm{j}\omega m}\,\mathrm{d}\omega\\ &\quad+\frac{1}{2\pi}\int_{-\pi}^{\pi}\log[\log Y(\omega)]'\mathrm{e}^{\mathrm{j}\omega m}\,\mathrm{d}\omega\\ &=c(m)+\mathscr{F}^{-1}\{\log\mathscr{F}[-\mathrm{j}mc(m)]\}.\end{aligned} \tag{22}$$

It can be seen that the $\eta(m)$ is the superposition of two terms. The first one is the cepstral coefficient $c(m)$, and the second one is the Fourier transform of the linearly liftered cepstral coefficients followed by a logarithm and an inverse Fourier transform. The latter one can be treated as nonlinearly liftered cepstral coefficients. While the linear liftering does not alter the performance, the effectiveness of this nonlinear liftering transform for HMM-based speech recognition will be shown through experimental results described later in this paper.

2.4. Comparison with the spectral subtraction

The spectral subtraction (SS) technique has been popularly employed to eliminate the detrimental effect of noise in robust speech recognition and speech enhancement. This technique is used to restore the spectrum of clean speech signal by subtracting an estimated noise spectrum from that of the noisy signal in the frequency domain, thereby yielding noise-free spectrum. Recalling the power spectra of noisy speech, noiseless speech and the noisy signal defined in (2) and (4), SS can be formulated as (Boll, 1979; Nolazco Flores and Young, 1994):

$$\widehat{X}(\omega)=\begin{cases}Y(\omega)-\alpha\widehat{V}(\omega), & \text{if } Y(\omega)>\frac{\alpha}{1-\beta}\widehat{V}(\omega),\\ \beta\widehat{V}(\omega), & \text{otherwise,}\end{cases} \tag{23}$$

where $\widehat{X}(\omega)$ and $\widehat{V}(\omega)$ are speech signal estimate and noise estimate respectively, β defines the spectral flooring, and α controls the amount of noise subtracted from the noisy signal. For full noise subtraction, $\alpha=1$, for over-subtraction $\alpha>1$, and for under-subtraction, $\alpha<1$.

SS can be performed either in each frequency bin or on sub-band basis. For speech recognition, it was found that SS operated in each band-pass filter could yield more consistent improvement for MFCC features against noise (Chen et al., 2001). In brief, this SS is expressed as

$$\widehat{E}_X(k) = \begin{cases} E_Y(k) - \alpha\widehat{E}_V(k), & \text{if } E_Y(k) > \frac{\alpha}{1-\beta}\widehat{E}_V(k), \\ \beta\widehat{E}_V(k), & \text{otherwise,} \end{cases} \tag{24}$$

where $E_Y(k)$ is the output of the kth band-pass filter when $Y(k)$ is passed though the triangular filter bank, and $\widehat{E}_X(k)$ and $\widehat{E}_V(k)$ are estimated filter bank energies for clean speech and noise signal.

The central issue for such a SS scheme is to compute $\widehat{E}_V(k)$. In this paper, the $\widehat{E}_V(k)$ is obtained sequentially by

$$\widehat{E}_V(k) = \gamma\widehat{E}_V^-(k) + (1-\gamma)E_Y(k), \tag{25}$$

where $\widehat{E}_V^-(k)$ suggests the noise estimate in previous frame, and γ takes on different "attack" and "decay" values depending on the relationship of $E_Y(k)$ to the previous estimate, i.e.,

$$\gamma = \begin{cases} \gamma_a = 0.99, & \text{if } E_Y(k) > \widehat{E}_V^-(k), \\ \gamma_d = 0.90, & \text{if } E_Y(k) \leqslant \widehat{E}_V^-(k). \end{cases} \tag{26}$$

Recalling its definition given in (8), we can rewrite DPS as

$$\begin{aligned} D(k) &= \sum_{l=-O}^{P} b_l Y(k+l) \\ &= b_0 Y(k) - \sum_{\substack{l=-O \\ l\neq 0}}^{P} b_l Y(k+l). \end{aligned} \tag{27}$$

Comparing (27) with (23), one can see the difference between the DPS representation and the spectral subtraction technique. In this paper, the comparison between DPS and the spectral subtraction technique is made for noisy speech recognition. Results will be presented in the next section.

3. Experiments

The proposed feature has been extensively tested on many tasks, which include various operating environments. For brevity, we cite only some of them in this paper.

3.1. Isolated speech recognition

The first experiment uses the TI46 database to find out which form of DPS given in (9)–(11) can lead to a better recognition performance. TI46 is an isolated spoken words database which was 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 remaining 16 are designated as testing tokens. Speech signal is digitized at a sampling rate of 12.5 kHz 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, DPSCC1, DPSCC2, and DPSCC3.

MFCC: Speech signal is analyzed every 10 ms with a frame width of 32 ms (with preemphasis and Hamming windowing). For each frame, a 512-point FFT is performed to estimate its power spectrum, which is then fitted to a mel-frequency filter bank consisting of 24 triangular filters. 12 MFCCs are computed by applying a logarithm and a cosine transform to the 24 filter bank energies (the MFCC of order 0 is ignored).

DPSCC1: For each frame, the power spectrum is estimated. The differential power spectrum is then calculated according to (9). The magnitude of DPS is then input to a same mel-frequency filter bank and is converted to 12 cepstral

coefficients. Similarly, we compute the *DPSCC2* and *DPSCC3* according to (10) and (11) respectively.

The recognition system used is a multi-speaker 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 four 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. Experimental results are shown in Table 1.

From above results, we can make several observations. First, the DPS based features can at least yield comparable performance as the standard MFCCs. This indicates that, just as power spectrum, DPS preserves the information of speech signal necessary for speech recognition. Second, for both MFCCs and DPSCCs, the inclusion of dynamic and acceleration features can greatly augment the recognition performance. Third, among the three types of DPS definitions, DPS1 defined in (9) yields the best performance. It gives 23% word error rate reduction as compared to the MFCC baseline.

In the subsequent experiments, we will evaluate the DPS based cepstrum and its robustness with respect to noise for various tasks. As we have shown that DPSCC1 yields the most promising result, we will only assess the DPSCC1 based features. For brevity, we shall, from now on, drop the 1 from DPSCC1 without introducing any confusion.

3.2. SNR improvement

Before we go on further to conduct noisy speech recognition, let us first examine the frame based SNR for both power spectra and differential power spectra of a speech signal. In this experiment, the clean speech signals are taken from the speaker m1 in the TI46 database. We take Lynx noise from the NOISEX database (Varga et al., 1992). The noise signal is downsampled from 16 to 12.5 kHz to match the bandwidth of speech signal and is then added to control the utterance level SNR to a certain dB.

For power spectrum based representation, by assuming that the noise and speech signal are uncorrelated, we introduce frame level SNR which is defined as

$$\mathrm{SNR}_Y(i) = 10\log_{10}\frac{\sum_{k=0}^{K/2} X(i,k)}{\sum_{k=0}^{K/2} V(i,k)}, \tag{28}$$

where $X(i,k)$, $V(i,k)$ are power spectra for the ith frame of noiseless speech and noise signal defined in (4), and K is the length of FFT.

Similarly, we can define the frame SNR for DPS if the same uncorrelation assumption is made,

$$\mathrm{SNR}_D(i) = 10\log_{10}\frac{\sum_{k=0}^{K/2} D_X(i,k)}{\sum_{k=0}^{K/2} D_V(i,k)}, \tag{29}$$

where $D_X(i,k)$, $D_V(i,k)$ are differential power spectra defined in (7).

Having defined the frame level SNR, we now start to compare $\mathrm{SNR}_Y(i)$ and $\mathrm{SNR}_D(i)$ for noisy speech signal. Fig. 4(a) shows an '*a*' sound. We add some Lynx noise to control the utterance level SNR to be 10 dB. The noisy speech signal is shown in Fig. 4(b). Fig. 4(c) plots both SNR_Y and SNR_D as a function of frame index i. One can see, from the plot, that for speech part, SNR_D is about 2–7 dB higher than SNR_Y. Fig. 5 plots a similar graph but for an '*i*' sound. Again we found that for the voiced part, SNR_D is higher than SNR_Y.

Table 1
Word accuracy (%) using different feature sets

	MFCC	DPSCC1	DPSCC2	DPSCC3
12 S	84.1	86.1	85.1	85.7
12 S + 12 D	90.4	92.1	90.9	92.1
12 S + 12 D + 12 A	91.8	93.6	91.7	92.8

S: static features; D: dynamic features which are calculated by subtracting the two preceding from the two following static feature vectors; A: accelerations which are calculated by subtracting the two preceding from the two following dynamic feature vectors.

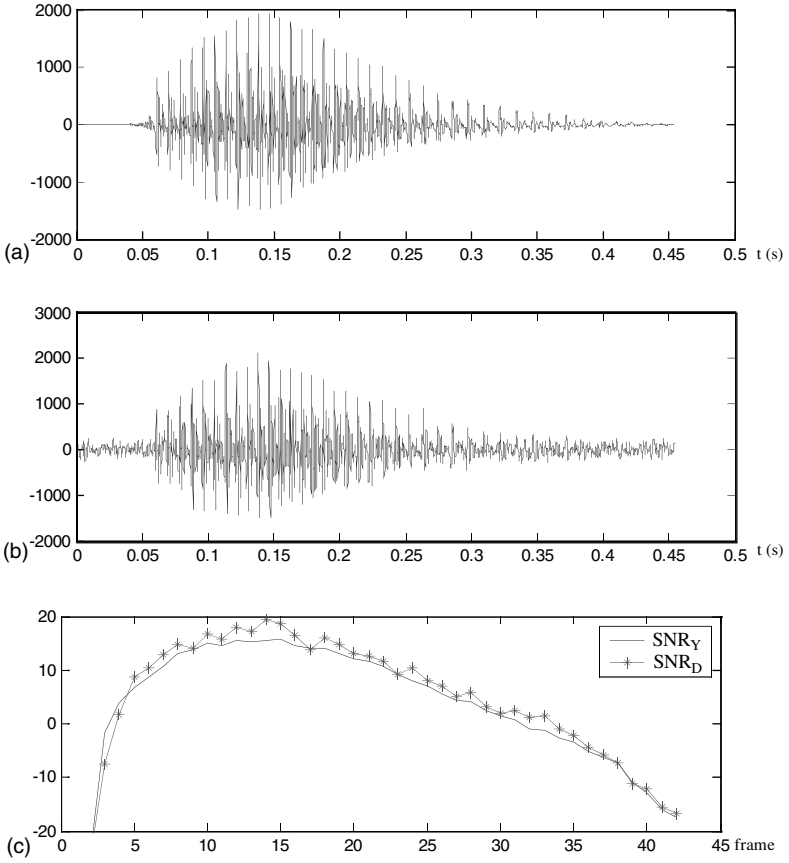

Fig. 4. Waveform plots of the clean and noisy signals and the frame level SNRs of the noisy signal of an 'a' sound: (a) Waveform plot of the clean 'a' sound; (b) Waveform plot of the 'a' sound in Lynx noise condition (SNR = 10 dB); (c) SNR_D and SNR_Y in Lynx noise environment.

We further examined many other sounds and various types of noise in 10 dB condition. We found that for voiced sound, the average SNR_D is approximately 4 dB higher than SNR_Y. For unvoiced speech and silence, SNR_D and SNR_Y are quite similar. For this reason, we can expect the DPS based feature to be more resilient to noise than the power spectrum based features, in other words, DPSCC should be more resilient to noise than MFCC.

3.3. Connected digits recognition

This experiment is to recognize connected digits. The TI connected digits database (Zue et al., 1990) is used for this purpose. This database contains digit strings uttered by adult and child speakers. However, only digit strings from 225 adult speakers are used in this experiment. These strings are originally divided into a training set and a test set for consistency in comparison of results among different researchers.

The vocabulary in this database consists of 11 words which include 10 digits and an "oh". Each speaker 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 of 20 kHz. We downsampled speech to 8 kHz using the Matlab downsampling function.

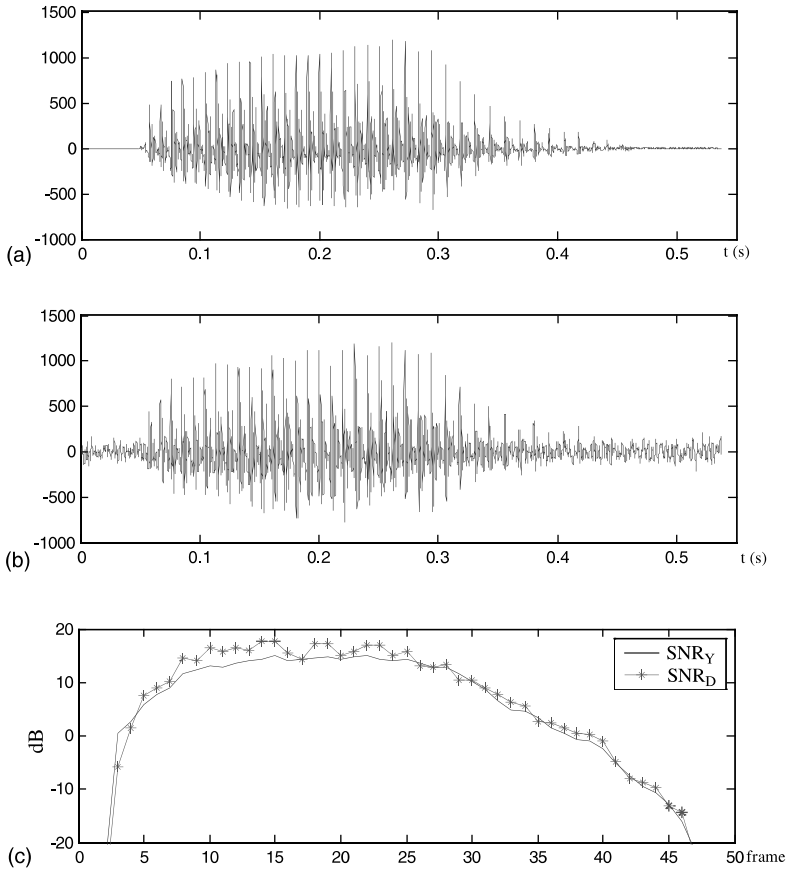

Fig. 5. Waveform plots of the clean and noisy signals and the frame level SNRs of the noisy signal of an '*i*' sound: (a) Waveform plot of the clean '*i*' sound; (b) Waveform plot of the '*i*' sound in Lynx noise condition (SNR = 10 dB); (c) SNR_D and SNR_Y in Lynx noise environment.

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. Noise signal is taken from the NOISEX database (Varga et al., 1992). It consists of various types of noise signals, among which three representative types of noise are selected to report here. They are wide-band stationary speech noise, narrow-band stationary Lynx helicopter noise, and nonstationary machine-gun noise. We refer the reader to (Varga et al., 1992) for detailed descriptions of the characteristics of these noises. The noise signal provided in this database is sampled at 16 kHz. 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 speaker-independent mixture Gaussian HMM system. The model set consists of 11 word-models, a silence model and a short pause model. With the exception of 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 15 ms with a frame width of 32 ms (with preem-

phasis and Hamming windowing). Each frame is transformed into 12 MFCCs using the same procedure as that in Experiment 1. Moreover, the normalized logarithmic short-time energy is also added to 12 MFCCs to form a 13-dimensional static vector. This static vector is then expanded to produce a 39-dimensional feature vector (static + delta + acceleration).

DPSCC: Speech signal is split into frames as described above. For each frame, the power spectrum is estimated and the differential power spectrum is then calculated according to (9). The magnitude of DPS is converted to 12 cepstral coefficients. This 12-dimensional vector is further expanded to a 39-dimensional feature vector using same strategy as used to compute the MFCC features.

MFCC + CMN: MFCC features with cepstral mean normalization (CMN).

DPSCC + CMN: DPSCC features with CMN.

The experiment results are shown in Fig. 6.

From Fig. 6, we can make following observations:

(1) As compared with the conventional MFCCs, the new cepstral vector derived from DPS yields at least comparable performance in clean, as well as high, SNR conditions.
(2) In most strong noise conditions, DPSCC outperforms MFCC.
(3) CMN is effective to augment the robustness of both MFCC and DPSCC with respect to noise.
(4) After CMN, the DPS features outperform MFCC in both clean and noisy conditions.

3.4. Phone recognition

The fourth experiment is to perform phone recognition. The speech data employed in this experiment is the TIMIT phoneme based continuous speech database (Lamel et al., 1986), 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 16 kHz with 16 bits per-word.

The TIMIT database is phonetically transcribed using a set of 61 phones. To facilitate comparison with the results reported in the literature (Lee and Hon, 1989), we perform phonetic recognition on this database over the set of 39 classes that are commonly used for such evaluation. Again, the HTK toolkit is configured to perform the recognition task. The model set consists of 39 monophone HMMs. Each model has three emitting states. An eight-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 + CMN* (39 coefficients) and *DPSCC + CMN* (39 coefficients). The static MFCCs and DPS based cepstral coefficients are estimated using the same procedure as described in the previous experiment. The 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 Fig. 7.

We can make the following observations from this figure:

(1) The MFCC and the DPSCC features yield comparable results in clean and weak noise conditions.
(2) DPSCC features slightly outperform the MFCC features in strong noise conditions.

3.5. Evaluation on AURORA task

The AURORA task (Hirsch, 2000) has been defined by the European Telecommunications Standards (ETSI) as a cellular industry initiative 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 20–8 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

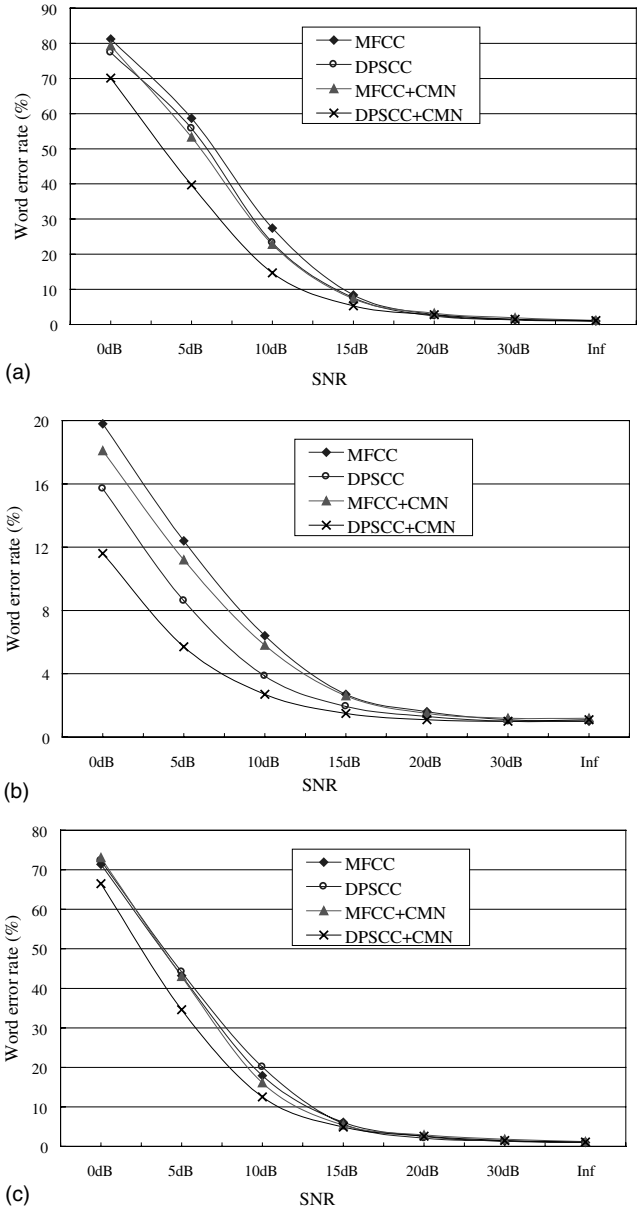

Fig. 6. Performance in different noise conditions: (a) word error rate in the wide-band stationary speech noise conditions; (b) word error rate in the machine-gun noise conditions; (c) word error rate in the Lynx noise conditions. (SNR = Inf means no noise is added to the original signal.)

additional filtering is applied. The two "standard" frequency characteristics used are G.712 and MIRS (Hirsch, 2000). Noise is artificially added to the filtered TIDigits at SNRs of 20, 15, 10, 5, 0 and −5 dB. Noise signals are recorded at different places including suburban train, crowd of people (babble), car, exhibition hall, restaurant, street, airport and train station.

Two training modes are defined, i.e., training on clean data only and training on clean as well as

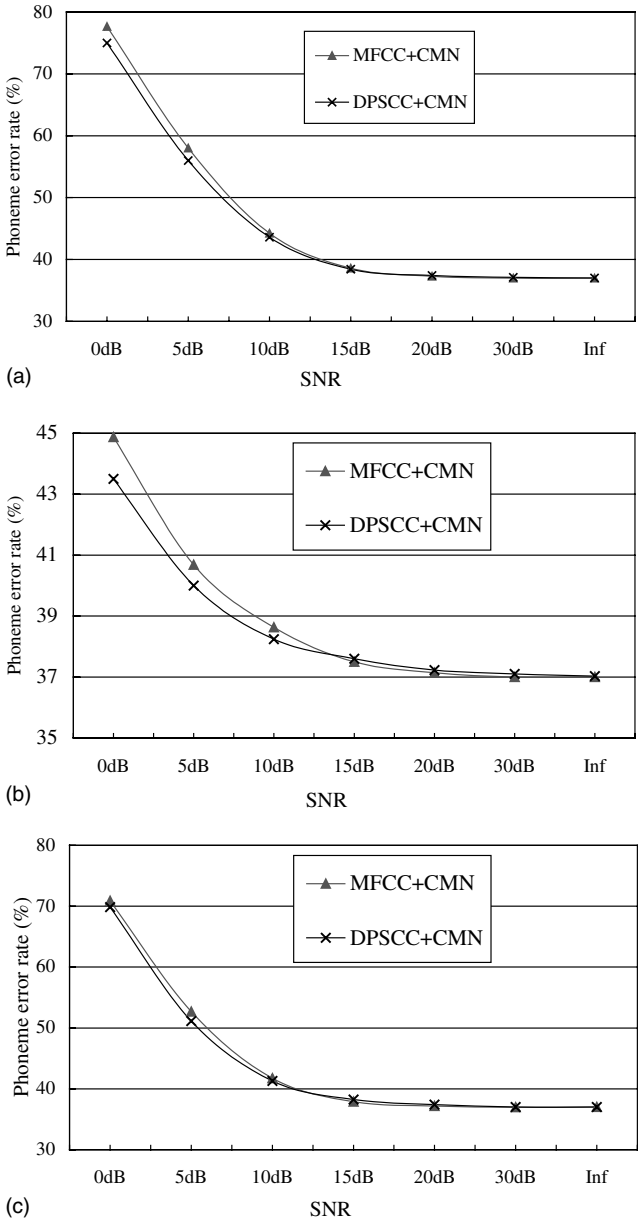

Fig. 7. Phoneme recognition performance in different noise conditions: (a) phoneme error rate in the wide-band stationary speech noise conditions; (b) phoneme error rate in the machine-gun noise conditions; (c) phoneme error rate in the Lynx noise conditions.

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 TI-Digits training parts 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, 20, 15, 10, 5 dB and the

clean condition. Both speech and noise are filtered before adding together.

Three test sets are defined. 4004 utterances from 52 male and 52 female speakers in the TIDigits test part are divided into four 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 20 to −5 dB in decreasing steps of 5 dB after speech and noise are being filtered with the G. 712. The three subsets are as below:

Test Set A: Suburban train, babble, car and exhibition noises are added to the four 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 the same noises as used for the multi-condition training mode.

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

Test Set C: It contains two of the four 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 20, 15, 10, 5, 0, and −5 dB.

To facilitate comparison of results among different researchers (Zhu et al., 2001; Kotnik et al., 2001; Andrassy et al., 2001; Droppo et al., 2001), AURORA task provides a reference recognizer which is based on the HTK software package. The model 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 a single state.

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

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

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

We assess the proposed feature on this task. DPSCCs are computed in exactly the same way except that our mel-frequency 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 average word recognition accuracies for three test sets in different noisy conditions are presented in Table 2. From the results, we can make the following observations:

(1) With the use of CMN, the average word error rate is reduced 8.8%. This shows the effectiveness of the CMN on robust speech recognition.
(2) SS is effective in dealing with additive noise. Used together with the CMN, it increases the average performance by 19.3%.
(3) The DPS based cepstrum outperforms MFCC. It also yields a slightly better performance than SS.

4. Discussion and conclusion

The concept of the differential power spectrum (DPS) was introduced and a new set of cepstral features was proposed in this paper for improving the robustness of speech recognition. We note that just like the power spectrum, DPS can also preserve spectral information to discriminate among different linguistic units (e.g., phonemes and words). Based on the analysis of the frame level SNR, we found that DPS had a higher SNR than the power spectrum, specially for voiced frames. This suggests that the DPS based features should be more resilient to noise than the power spectrum based features.

Compared with the linear liftering technique, the DPS based features can be decomposed as the superposition of conventional cepstral coefficients and their nonlinearly liftered counterpart. While a

Table 2
Recognition performance for different feature sets

	A				B				C		Overall ave
	Sub-way	Bab-ble	Car	Exhib.	Rest.	Street	Air-port	Sta-tion	Sub.	Street	
Average word accuracy (%)											
MFCC	89.1	88.4	86.8	88.0	86.6	87.8	88.3	86.2	83.5	85.7	87.0
MFCC + CMN	90.1	88.4	87.2	87.9	86.4	87.9	88.9	86.0	87.3	85.7	87.6
MFCC + SS + CMN	90.4	89.0	89.7	88.8	88.2	88.4	91.0	87.4	89.5	87.8	89.0
DPSCC + CMN	91.5	89.2	89.0	89.2	87.6	89.4	90.3	87.5	90.7	88.8	89.3
Relative error rate reduction (%) in comparison with the baseline MFCC											
MFCC	–	–	–	–	–	–	–	–	–	–	–
MFCC + CMN	11.6	3.9	4.9	−1.1	6.9	6.8	10.5	6.8	24.1	8.9	8.8
MFCC + SS + CMN	14.89	8.7	23.4	6.4	19.5	10.6	27.3	15.7	37.2	22.0	19.3
DPSCC + CMN	24.6	10.7	18.4	9.6	15.26	18.5	21.8	16.8	44.6	28.5	21.6

Recognition is performed for various conditions including clean speech and SNRs at 20, 15, 10, 5, 0 and −5 dB. According to the AURORA standard, the average in each type of noise is computed by averaging the word accuracy in 20, 15, 10, 5, and 0 dB, while both clean and −5 dB conditions are ignored.

linear liftering transform was shown to have no effect on continuous density HMM-based speech recognition, the DPS based cepstrum can increase recognition performance in various noise conditions. The proposed features were also found to outperform the spectral subtraction technique.

The DPSCC feature was extensively evaluated on many recognition tasks, which include various operating environments. Results reveal that the proposed feature can yield at least comparable performance when compared to the conventional MFCCs. In most cases, it outperforms MFCC.

Compared to the estimation of MFCC, the extraction of DPSCC only requires $(K/2-1)$ more addition (subtraction) and absolute operations for each frame signal (where K is the length of FFT). This increase in computational complexity is negligible for today's computer.

References

Andrassy, B., Vlaj, D., Beaugeant, C., 2001. Recognition performance of the Siemens front-end with and without frame dropping on the AURORA 2 database. Proc. EUROSPEECH, Scandinavia, pp. 193–196.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics, Speech Signal Process. 27 (2), 113–120.

Bourlard, H., Dupont, S., 1996. A new ASR approach based on independent processing and recombination of partial frequency bands. Proc. ICSLP, Philadelphia, pp. 426–429.

Chen, J., Paliwal, K.K., Nakamura, S., 2001. Sub-band based additive noise removal for robust speech recognition. Proc. EUROSPEECH, Scandinavia, pp. 571–574.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech Signal Process. 28, 357–366.

Droppo, J., Deng, L., Acero, A., 2001. Evaluation of the SPLICE algorithm on the Aurora2 database. Proc. EUROSPEECH, Scandinavia, pp. 217–220.

Furui, S., 1986. Speaker-independent isolated work recognition using dynamic features of speech spectrum. IEEE Trans. Acoustics, Speech Signal Process. 34 (1), 52–89.

Gales, M.J.F., Young, S.J., 1996. Robust speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4 (5), 352–359.

Geller, D., Haeb-Umbach, R., Ney, H., 1992. Improvements in speech recognition for voice dialing in the car environment. Proc. ESCA Workshop on Speech Processing in Adverse Conditions, Mandelieu, pp. 203–206.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoustic. Soc. Am. 87 (4), 1738–1752.

Hermansky, H., Morgan, N., 1994. RASTRA of processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Hermansky, H., Morgan, N., Bayya, A., Kohn, P. 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). Proc. EUROSPEECH, Genova, pp. 1367–1370.

Hirsch, H., Meyer, P., Ruehl, H., 1991. Improved speech recognition using high-pass filtering of subband envelopes. Proc. EUROSPEECH, Genova, pp. 413–416.

Hirsch, H.G., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ISCA ASR2000, Paris, France.

Juang, B.H., Rabiner, L.R., Wilpon, J.G., 1987. On the use of bandpass liftering in speech recognition. IEEE Trans. Acoust., Speech Signal Process. 35 (7), 947–954.

Junqua, J.-C., Cerisara, C., Rigazio, L., Kryze, D., 2001. Environment-adaptive algorithms for robust speech recognition. Proc. Internat. Workshop Handsfree Speech Communication, Kyoto, Japan, pp. 31–34.

Kim, D.-S., Lee, S.-Y., Kil, R.M., 1999. Auditory processing of speech signals for robustness speech recognition in real-world noisy environments. IEEE Trans. Speech Audio Process. 7 (1), 55–69.

Kotnik, B., Kacic, Z., Horvat, B., 2001. A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm. Proc. EUROSPEECH, Scandinavia, pp. 197–200.

Lamel, L.F., Kassel, H.K., Seneft, S., 1986. Speech database development: Design and analysis of the acoustic–phonetic corpus. Proc. DARPA Speech Recognition Workshop, Palo Alto, pp. 100–109.

Lee, K.-F., Hon, H.-W., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoustics, Speech Signal Process. 37 (11), 1641–1648.

Moreno, P.J., Raj, B., Stern, R.M., 1996. A vector Taylor series approach for environment independent speech recognition. Proc. ICSLP, Philadelphia, PA, pp. 733–736.

Nolazco Flores, J.A., Young, S.J., 1994. Continuous speech recognition in noise using spectral subtraction and HMM adaptation. Proc. ICASSP, Adelaide, Australia, pp. 409–412.

Ohkura, K., Sugiyama, M., Sagayama, S., 1992. Speaker adaptation based on transfer vector field smoothing technique. Proc. ICSLP, Banff, Canada, pp. 369–372.

Paliwal, K.K., 1992. On the performance of the frequency-weighted cepstral coefficients in vowel recognition. Speech Commun. 18, 151–154.

Paliwal, K.K., 1999. Decorrelated and liftered filter-bank energies for robust speech recognition. Proc. EUROPSEECH, Budapest, pp. 85–88.

Picone, J.W., 1993. Signal modeling techniques in speech recognition. Proc. IEEE 81 (9), 1215–1247.

Popescu, D.C., Zeljkovic, I., 1998. Kalman filtering of colored noise for speech enhancement. Proc. ICASSP, Seattle, pp. 997–1000.

Rahim, M., Juang, B.-H., 1996. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. IEEE Trans. Speech Audio Process. 4 (1), 19–30.

Sagayama, S., Yamaguchi, Y., Tahahashi, S., Takahashi, J.-I., 1997. Jacobian approach to fast acoustic model adaptation, Proc. ICASSP, Munich, Germany, pp. 835–838.

Soong, F.K., Rosenberg, A.E., 1986. On the use of instantaneous and transitional spectral information in speaker recognition. Proc. ICASSP, Tokyo, Japan, pp. 877–880.

Tohkura, Y., 1987. A weighted cepstral distance measure for speech recognition. IEEE Trans. Acoust., Speech Signal Process. 35 (10), 1414–1422.

Varga, A., Steeneken, H.J.M., Tomlinson, M., Jones D., 1992. The NOISEX-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.

Vaseghi, S.V., Milner, B.P., 1997. Noise compensation methods for hidden Markov model speech recognition in adverse environments. IEEE Trans. Speech Audio Process. 5 (1), 11–21.

Woodland, P.C., Gales, M.J.E., Pye, D., 1996. Improving environmental robustness in large vocabulary speech recognition. Proc. ICASSP, Atlanta, GA, pp. 65–68.

Zhu, Q., Iseli, M., Cui, X., Alwan, A., 2001. Noise robust feature extraction for ASR using Aurora 2 database. Proc. EUROSPEECH, Scandinavia, pp. 185–188.

Zue, V., Seneff, S., Glass, J., 1990. Speech database development at MIT: TIMIT and beyond. Speech Commun. 9, 351–356.