Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 48 (2006) 1294–1307
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition

Zekeriya Tufekci *, John N. Gowdy, Sabri Gurbuz [1], Eric Patterson [2]

105 Riggs Hall, Clemson University, ECE Department, Box 340915, Clemson, SC 29634-0915, USA

Received 19 December 2005; received in revised form 22 June 2006; accepted 22 June 2006

Abstract

Interfering noise severely degrades the performance of a speech recognition system. The Parallel Model Compensation (PMC) technique is one of the most efficient techniques for dealing with such noise. Another approach is to use features local in the frequency domain, such as Mel-Frequency Discrete Wavelet Coefficients (MFDWCs). In this paper, we investigate the use of PMC and MFDWC features to take advantage of both noise compensation and local features (MFDWCs) to decrease the effect of noise on recognition performance. We also introduce a practical weighting technique based on the noise level of each coefficient. We evaluate the performance of several wavelet-schemes using the NOISEX-92 database for various noise types and noise levels. Finally, we compare the performance of these versus Mel-Frequency Cepstral Coefficients (MFCCs), both using PMC. Experimental results show significant performance improvements for MFDWCs versus MFCCs, particularly after compensating the HMMs using the PMC technique. The best feature vector among the six MFDWCs we tried gave 13.72 and 5.29 points performance improvement, on the average, over MFCCs for −6 and 0 dB SNR, respectively. This corresponds to 39.9% and 62.8% error reductions, respectively. Weighting the partial score of each coefficient based on the noise level further improves the performance. The average error rates for the best MFDWCs dropped from 19.57% to 16.71% and from 3.14% to 2.14% for −6 dB and 0 dB noise levels, respectively, using the weighting scheme. These improvements correspond to 14.6% and 31.8% error reductions for −6 dB and 0 dB noise levels, respectively.

Keywords: Noise robust ASR; Wavelet; Local feature; Feature weighting

1. Introduction

Real world applications require that speech recognition systems be robust to interfering noise. Unfortunately, though, the performance of a speech recognition system drops dramatically when there is a mismatch between training and testing conditions. Many different approaches have been studied to decrease the effect of noise on the recognition

* Corresponding author. Present address: Izmir Yuksek Teknoloji Enstitusu, Elektrik-Elektronik Muhendisligi Bolumu, 35430 Urla-izmir, Turkey. Tel.: +90 232 750 6288; fax: +90 232 750 6505.

E-mail addresses: zekeriyatufekci@iyte.edu.tr (Z. Tufekci), john.gowdy@ces.clemson.edu (J.N. Gowdy), sabrig@his.atr.jp (S. Gurbuz), pattersone@uncw.edu (E. Patterson).

[1] Currently at NICT/ATR Cognitive Information Science Laboratories, Japan.

[2] Currently at University of North Carolina at Wilmington, Department of Computer Science, Wilmington, NC 28403, USA.

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.06.006

performance (Gong, 1995). These approaches can be divided into three main groups: speech enhancement techniques (Boll, 1979; Cung and Normandin, 1992; Lockwood and Boudy, 1991), inherently robust speech features (Mansour and Juang, 1989; Ghitza, 1986), and model-based compensation techniques (Beattie and Young, 1991, 1992; Berstein and Shallom, 1991; Klatt, 1979; Mellor and Varga, 1992; Varga and Moore, 1990; Gales and Young, 1992, 1993a, 1995b, 1996).

One of the most effective and popular model-based techniques for dealing with noisy speech is Parallel Model Compensation (Gales and Young, 1992, 1993a, 1995b, 1996). This technique attempts to estimate models of noisy given clean speech models and a noise model. Previous work (Gales and Young, 1995a) has produced results near the optimal (matched) condition (when training and testing noise environments are identical).

In addition to the methods mentioned above, recognition systems based on features local in the frequency domain, such as multiband (Bourlard and Dupont, 1996; Hermansky et al., 1996; Tufekci and Gowdy, 2001) and multiresolution (Vaseghi et al., 1997; Chengalvarayan, 1999; Tufekci and Gowdy, 2000; Gowdy and Tufekci, 2000) techniques, have received great attention for dealing with noisy speech. In this paper, the speech feature vector will be referred to as local if some of the coefficients of the vector represent local information in the frequency domain, even though the other coefficients do not.

Conventional feature extraction methods use the entire frequency band to extract speech features for speech recognition. However, as pointed out by Fletcher (1953) (and reviewed by Allen (1994)), the Human Speech Recognition (HSR) system seems to utilize partial recognition information across frequencies, probably in the form of speech features that are local in frequency. Fletcher's work (Fletcher, 1953) led to the subband-based speech recognizer (Bourlard and Dupont, 1996; Hermansky et al., 1996). Hermansky et al. (1996) and Bourlard and Dupont (1996) also proposed subband-based speech recognition systems. They simply divided the frequency band into subbands, extracted features for each subband, and then calculated scores for each subband. Finally, they combined each subband's recognition score by using merging techniques.

There are three main motivations for local (in frequency domain) feature-based recognizers:

- Some subbands of the speech spectrum are inherently more relevant than others for the task of speech recognition. Therefore, the contribution of each subband to the overall recognition decision can be weighted depending on the information that each subband conveys.
- Transitions between more stationary segments of speech do not necessarily occur at the same time across the different frequency bands. The local feature-based approach may have the potential of relaxing the synchrony inherent in current HMM systems. A frame of speech may contain information of two adjacent phonemes. If one of these phonemes is voiced and the other is unvoiced, then the low-frequency spectrum is dominated by voiced-phoneme information, and the high-frequency spectrum is dominated by the unvoiced-phoneme information. In traditional feature extraction methods that are based on extracting speech features using the full frequency band, we inherently assume that a speech frame conveys information about only one phoneme at a time. However, this is not always true. Frequency-local features may help model the possible asynchrony more accurately.
- Local features are affected differently in noisy environments. When one frequency band is corrupted by noise, only a few coefficients are affected if the coefficients represent local information; otherwise, the noise affects all coefficients. Even if the whole frequency band is corrupted by noise, the SNR will likely be different for each subband. Local features allow a SNR-weighted contribution of each coefficient to the global score.

Previous work on a phoneme-recognition task showed that local features (Tufekci and Gowdy, 2000, 2001; Gowdy and Tufekci, 2000) outperform MFCCs for recognition of clean speech. This indicates that some subbands of the speech spectrum are inherently more relevant than others to the task of speech recognition. Asynchrony between frequency bands has been studied (Mirghafori and Morgan, 1998; Cerisara et al., 1998; Tomlinson et al., 1997; Bourlard and Dupont, 1996), and it has been shown that accommodating asynchrony between frequency bands also improves performance.

The subband-based recognizer (Bourlard and Dupont, 1996; Hermansky et al., 1996) is one of the most popular recognizers based on local features. Linear Predictive Cepstral Coefficients

(LPCCs) or MFCCs (Davis and Mermelstein, 1980) for each subband are typically used as subband features. Therefore, the resulting features are a Cosine Transform (CT) of a preprocessed log magnitude spectrum of subbands or Discrete Cosine Transform (DCT) of mel-scaled log filterbank energies of subbands. Subband based recognizers have two drawbacks:

- The length and shape of basis vectors determine the resolution capability of a transformation (Mallat, 1998). For longer basis vectors the transformation will have better frequency resolution but worse time resolution. The basis vector that is well concentrated (because of its shape) in time and frequency domain will have better resolution in time and frequency than the basis vector that is not well concentrated in time and frequency for a given basis vector length.
  In this paper, the DCT is applied to the mel-scaled log-filterbank energies of a speech frame. The basis vectors of the DCT (CT) have approximately the same resolution in time and frequency since same length windows are used in calculating the cepstral coefficients. That is, the frequency resolution of vector-1 is approximately equal to the frequency resolution of vector-2,...,vector-N and the time resolution of vector-1 is approximately equal to the time resolution of vector-2,...,vector-N. However, basis vectors with different time–frequency resolutions capability are needed to capture the changes in time and frequency.
- Basis vectors of the DCT also have the problem that high frequency artifacts (Mallat, 1998) can be introduced due to abrupt changes at window boundaries.

To overcome the former problem, Vaseghi et al. (1997) suggested multi-resolution features. However, since they use the same basis vectors as the DCT, the latter problem was still present. In our previous work (Tufekci and Gowdy, 2000; Gowdy and Tufekci, 2000), we proposed the use of DWT which has good time and frequency resolution, instead of the DCT, to solve the problems mentioned above. Also based on mel-frequency scaled bands, the resulting features are called MFDWCs. One important property of the DWT is that the inverse DWT exists, which is necessary for the PMC technique. Our previous work (Tufekci and Gowdy, 2000; Gowdy and Tufekci, 2000) has shown that MFDWCs outperform subband features, multiresolution features and MFCCs for clean speech and also for noisy speech.

It is known that the recognizer is optimal (Gales and Young, 1996) if the training and testing conditions are identical. A practical method for approaching the optimal recognizer for different noise conditions is PMC. This technique allows for estimation of HMM parameters in new environments. Since the features involved are local, the estimated HMM parameters for the new environment will represent local information. This is very important because the estimated parameters for a particular coefficient will be affected only if that particular coefficient is corrupted by noise.

In this work the PMC technique was implemented along with MFDWCs to test the performance of MFDWCs for noisy conditions. A proposed weighting method based on the noise level was also tested. The paper is organized as follows. The wavelet transform and MFDWCs are introduced in Section 2. Application of PMC to MFDWCs is explained in Section 3. Weighting the partial score of each coefficient based on the noise level is explained in Section 4. Section 5 gives the experimental setup and results, and conclusions are presented in Section 6.

2. Wavelet transform and MFDWCs

MFDWCs are obtained by applying the DWT to the mel-scaled log-filterbank energies of a speech frame. For information about the Wavelet Transform (WT) and implementations of the WT, interested readers may refer to Mallat (1998) and Vetterli and Kovacevic (1995). The WT uses short basis functions to measure the high frequency content of the signal and long basis functions to measure the low frequency content of the signal. This property of the WT distinguishes it from the Short Time Fourier Transform (STFT) and the Fourier Transform (FT). For an example, consider a field with different sizes of rocks as shown in Fig. 1. Consider that we want to know information about the location and size of the rocks, and assume that we have three different transformations to obtain information about the location and size of the rocks. The first one has resolution similar to the FT, which uses the entire field to obtain information. The second one has resolution similar to the STFT, which divides the field into same size subfields using rectangular windows and extracts

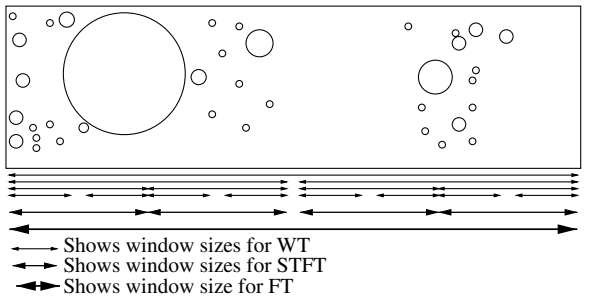

Fig. 1. A field with different sizes of rocks, illustrating differences between the FT, STFT, and WT.

features from each subfield. The third one has resolution similar to the WT, which divides the field into subfields using different sizes of windows that may overlap, and extracts features from each subfield. Fig. 1 illustrates division of the field into subfields for the three different transformations mentioned above.

If we use a transform which has resolution similar to the FT, we get information only about the number of rocks for each size but not about the locations of the rocks. If we use a transform which has resolution similar to the STFT, we also get location information, but we do not get good location information for small rocks (since the window size is large). Also, the rock size may be larger than the window size, so that we lose information for large rocks. However, since the transform that has resolution properties similar to the WT uses small windows to obtain information about small rocks, and large windows (which may overlap) to obtain information about the large rocks, we have a more effective approach to determining size and location of the rocks. This informal discussion presents the main motivation for using the WT for feature extraction.

A wavelet is a function $\Psi(t) \in L^2(\mathbb{R})$ (space of square-integrable functions) of zero average and unit norm so that

$$\int_{-\infty}^{+\infty} \Psi(t)\mathrm{d}t = 0, \tag{1}$$

$$\|\Psi(t)\| = 1. \tag{2}$$

The analysis function of wavelet transform at scale s and translation u is given by

$$\Psi_{u,s}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-u}{s}\right). \tag{3}$$

The wavelet transform of a function $f(t) \in L^2(\mathbb{R})$ at the time u and scale s is given by

$$WF(u,s) = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\Psi^*\left(\frac{t-u}{s}\right)\mathrm{d}t = \int_{-\infty}^{+\infty} f(t)\Psi^*_{u,s}(t)\mathrm{d}t, \tag{4}$$

where * denotes complex conjugate. Theoretically, any function with zero mean and finite energy can be a wavelet. There are many criteria, though, by which to choose a wavelet. Since we cannot implement a wavelet of infinite duration, we need compactly supported wavelets for practical applications. Decay of the wavelet in the frequency and time domains is important. We want the wavelet to decay quickly in time and frequency in order to have good locality in both these domains. Filterbank-based wavelets can be implemented efficiently. Since our signals are of finite length, the wavelet coefficients will have unwanted large variations at the borders because of the associated discontinuities (Mallat, 1998). We can use folded wavelets that require symmetric or anti-symmetric wavelets such as the spline wavelet to decrease the effect of discontinuities at the borders, or we can use border wavelets. Because of considerations given above, the options for choosing a wavelet are limited. In addition we need to use the Discrete Wavelet Transform instead of the Continuous Wavelet Transform since our signal is discrete. Fig. 2 illustrates extraction of the MFCCs and MFDWCs. The first five steps are the same for both as shown in Fig. 2. Only the last step is different in that we take the Discrete Cosine Transform (DCT) of the mel-scaled log-filterbank energies to calculate MFCCs or the DWT of the log-filterbank energies to calculate MFDWCs. The first step is to divide the speech signal into blocks using overlapping smooth windows such as Hamming, Hanning, etc. The next step is to take

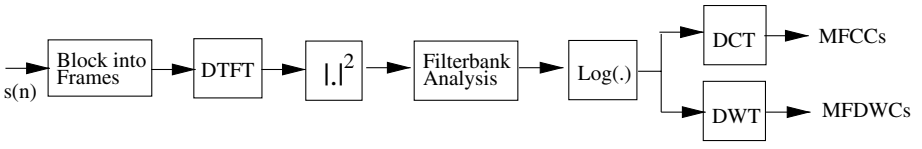

Fig. 2. Extraction of the MFCCs and MFDWCs.

the Discrete Time Fourier Transform (DTFT) of the windowed signal. Next, the square of the DTFT of the windowed signal is calculated. The outputs of the fourth step are the mel-scaled filterbank energies. The fifth step involves calculating the logarithm of the mel-scaled filterbank energies.

2.1. Localization property of the wavelet transform and cosine transform

In general, we take the transformation of a signal to get a more useful representation of the signal. However, this has different meanings depending upon the application. For coding, a goal is to represent the signal with fewer coefficients. For recognition, the objective is to separate the signals belonging to different categories in the new domain better than in the original domain. One of the important properties of certain linear transformations is localization which may lead to better representation when the signal is noisy. In this section, the localization property of a transform is explained. Let $\{\phi_k(t)\}$ be a set of functions on $L^2(\mathbb{R})$ (space of square-integrable functions) associated with transformation T. k is an index which may be multidimensional. Then define the transformation of $f(t)$ as

$$Tf(t) = F(k) = \int_{-\infty}^{+\infty} f(t)\phi_k^*(t)\mathrm{d}t = \langle f(t)\phi_k(t)\rangle, \tag{5}$$

where $\langle\ \rangle$ denotes inner product and * denotes complex conjugate. Parseval's theorem states that

$$Tf(t) = \int_{-\infty}^{+\infty} f(t)\phi_k^*(t)\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\Omega)\Phi_k^*(\Omega)\mathrm{d}\Omega, \tag{6}$$

where $F(\Omega)$ and $\Phi_k(\Omega)$ are the FT of $f(t)$ and $\phi_k(t)$, respectively. As seen from Eqs. (5) and (6) the transform of the signal depends on both $\phi_k(t)$ and the FT of $\phi_k(t)$, so the locality of $f(t)$ in the time and frequency domains depends on the spread of $\phi_k(t)$ in time and frequency, respectively. Therefore, it is desirable that $\phi_k(t)$ be well concentrated in the time and frequency domains to decrease the effect of noise. There are many ways to define the measure of locality of a function in the time and frequency domains. One approach (Mallat, 1998) uses of the variance of the function (σ_t^2) and the variance of the FT of the function (σ_Ω^2). σ_t^2 represents the spread of the function in time which is the measure of locality of the function in the time domain. The value σ_ω^2 represents the frequency spread of the function, which is the measure of locality of the function in the frequency domain.

The locality of a transformation of a signal is important in two ways for pattern recognition. First, different parts of the signal may convey different amount of information. When the coefficients represent local information, we can adjust the contribution of each coefficient to the total recognition rate depending on the information that each coefficient conveys. Second, when our signal is corrupted by noise that is local in time and/or in frequency, this noise affects only a few coefficients if our coefficients represent local information in time and frequency. Therefore, we can decrease the contribution of noise-corrupted coefficients to the overall recognition score depending on the SNR for noise corrupted coefficients.

Figs. 3–5 show the spread of wavelet basis functions in the time and frequency domains. All the wavelets shown in Figs. 3–5, which are labeled wavelet1 through wavelet6, were used in this study. Filterbank coefficients for these wavelets can be found in (Cohen et al., 1992). The cosine basis function is also included in the figures since the discrete version of the cosine transform is used to calculate the MFCC. The cosine basis functions are given as

$$\lambda_k\sqrt{\frac{2}{T}}\cos\left(\frac{\pi k t}{T}\right), \tag{7}$$

where $k \in \mathbb{N}$ (natural number). $\lambda_k = 1$ if $k \neq 0$ and $\lambda_k = 1/\sqrt{2}$ if $k = 0$. Fig. 3 shows the cosine basis function for $T = 2$ and $k = 4$.

As seen from Eqs. (5) and (6), if the basis functions are local, the resulting coefficients will represent local information. If the features represent local information, corruption of a frequency band will affect only a few coefficients which may lead to better performance. If we look at Figs. 3–5, it appears that wavelet1 is better concentrated in the frequency and time domains than the others. Theoretically, we expect a wavelet that is more concentrated in the frequency and time domains to perform better than the others. However, as will be shown in Section 5 where a comparison is given, this is not necessarily true. wavelet4 has the worst concentration in the frequency domain. Wavelet5 and wavelet6 have similar localization in the time and frequency domain, suggesting they may have similar performances. As shown in Fig. 6 the cosine basis function has discontinuities at the border

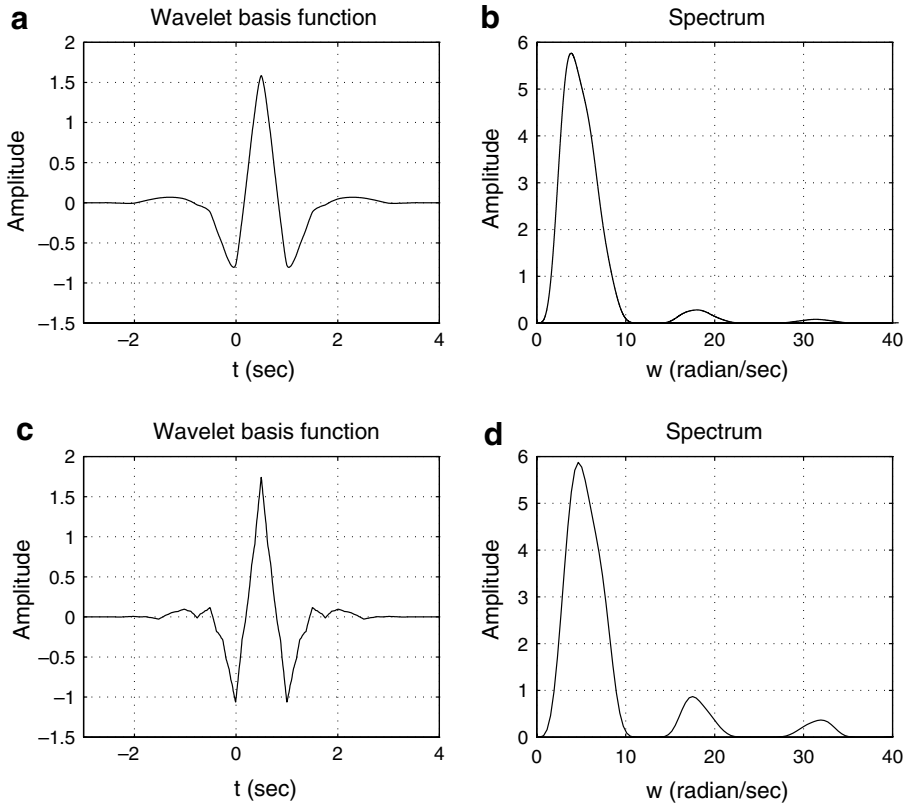

Fig. 3. Spreads of the basis functions in the time and frequency domains: (a) and (b) for wavelet1; (c) and (d) for wavelet2.

which may cause large variation (Mallat, 1998) at some coefficients. It is also not well concentrated in the frequency domain.

3. The PMC technique applied to the MFCCs and MFDWCs

In this section the PMC technique is briefly reviewed. Ideally, the training and testing conditions for a speech recognition system should be the same (this is the "matched system" condition). A logical approach to implementing such a system would be to retrain the system for each new test environment. However, it is often not practical to do this.

The PMC (Gales and Young, 1992, 1993a, 1995b, 1996) technique was proposed by Gales and Young to deal with new testing conditions by estimating the noisy speech model using a combination of clean speech and noise models. The PMC technique is very effective and less time consuming compared to retraining the system using the training data for a new environment. There are three PMC techniques for estimating the noisy speech parameters: numerical integration (Gales and Young, 1995b), a data-driven approach (Gales and Young, 1995a), and log-normal approximation (Gales and Young, 1992). We chose the log normal approximation approach since it demands the least computation for comparable performance.

When the noise is stationary, a single state noise model with one mixture may be sufficient to model it. When the noise is non-stationary or quasi-stationary, it may be necessary to use multiple mixtures (Yang and Haavisto, 1995) for the noise model. For example, the single mixture model may be sufficient for speech noise and Lynx helicopter noise. However, more mixtures are typically needed to model F16 and factory noise. STITEL noise is a periodic noise which may be better modeled by multi-state noise model (Gales and Young, 1993a). A multi-state noise model can be approximated with a single-state multiple mixture noise model. Therefore, in this paper, the noise was modeled by a single state with multiple mixture components. It is common practice to use delta coefficients to achieve better performance. Therefore, we also estimated the delta coefficients (Gales and Young, 1993b; Yang and Haavisto, 1996) using the PMC technique.

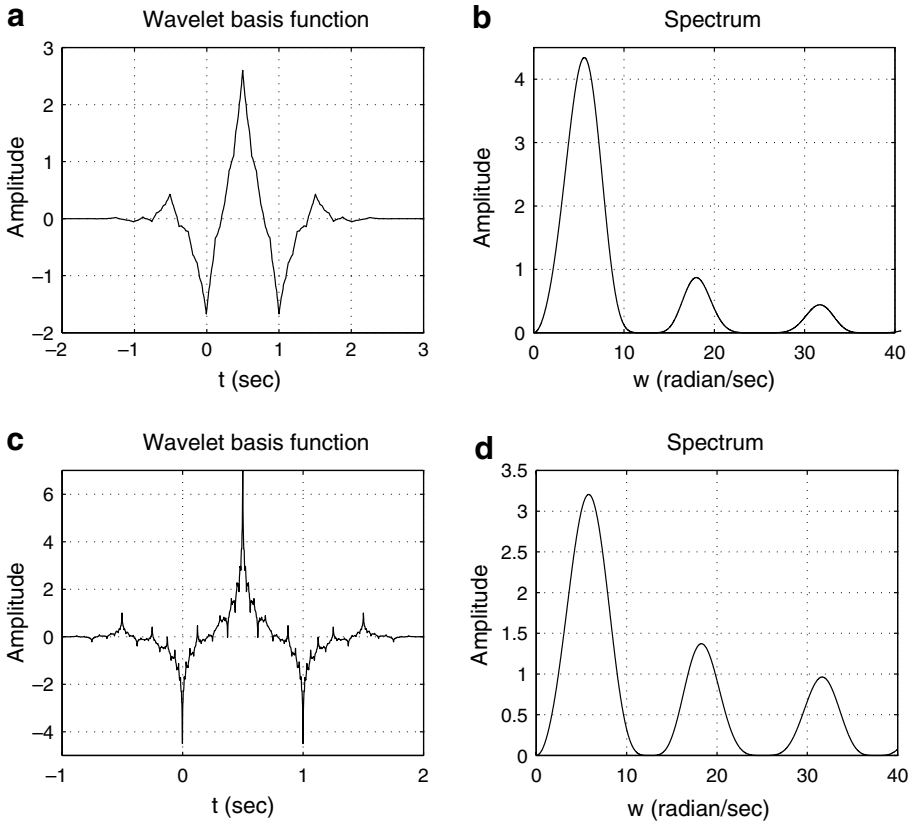

Fig. 4. Spreads of the basis functions in the time and frequency domains: (a) and (b) for wavelet3; (c) and (d) for wavelet4.

A series of assumptions, (Gales and Young, 1996) given below, are required to estimate the noisy speech parameters using the PMC technique:

- The speech and noise are independent.
- The speech and noise are additive in the time domain. In addition, it is assumed that there is sufficient smoothing of the spectral estimate so that speech and noise may be assumed to be additive at the power spectrum level.
- A multiple Gaussian mixture component model contains sufficient information to represent the distribution of the observation vectors in the cepstral or mel-scaled log filterbank energy domain.
- The frame/state alignment used to generate the speech models from the clean speech data is not altered by the addition of noise.

Three additional assumptions (Gales and Young, 1992, 1993b) are needed to use log normal approximation for estimating noisy-speech static and delta parameters as given below.

- The sum of two log-normally distributed random variables is approximately a log-normally distributed random variable.
- The variances of $\left(\frac{S_i}{S_1+N_i}\right)$ and $\left(\frac{N_i}{S_i+N_i}\right)$ are negligible where S_i and N_i are the ith components of the speech observation vector and noise observation vector, respectively, in the mel-scaled filterbank energy domain.
- $$E\left(\frac{S_i}{S_i+N_i}\right) \approx \left(\frac{\mu_i}{\mu_i+\tilde{\mu}_i}\right) = \gamma_i, \tag{8}$$
$$E\left(\frac{N_i}{S_i+N_i}\right) \approx \left(\frac{\tilde{\mu}_i}{\mu_i+\tilde{\mu}_i}\right) = \eta_i, \tag{9}$$
where E is expectation operator, μ_i and $\tilde{\mu}_i$ are the ith components of the clean speech and noise mean vectors in the mel-scaled filterbank energy domain.

For the rest of the paper, superscripts will be used to denote the domain of the observation or distribution, thus μ^c is the clean speech mean in the cepstral domain for the MFCCs or in the wavelet domain

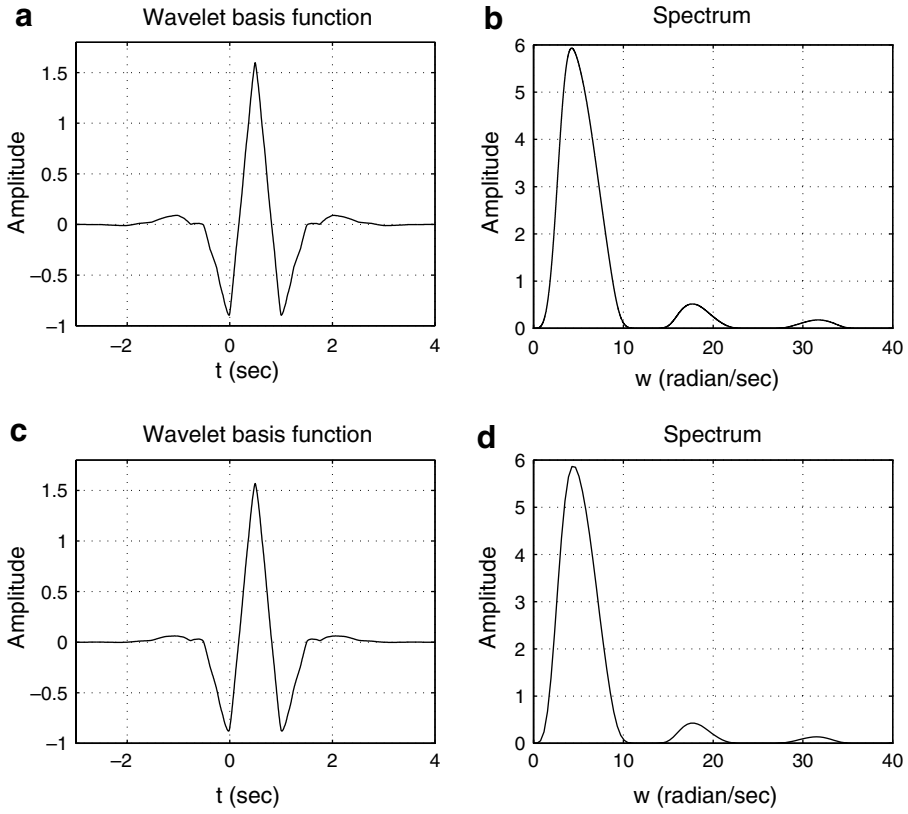

Fig. 5. Spreads of the basis functions in the time and frequency domains: (a) and (b) for wavelet5; (c) and (d) for wavelet6.

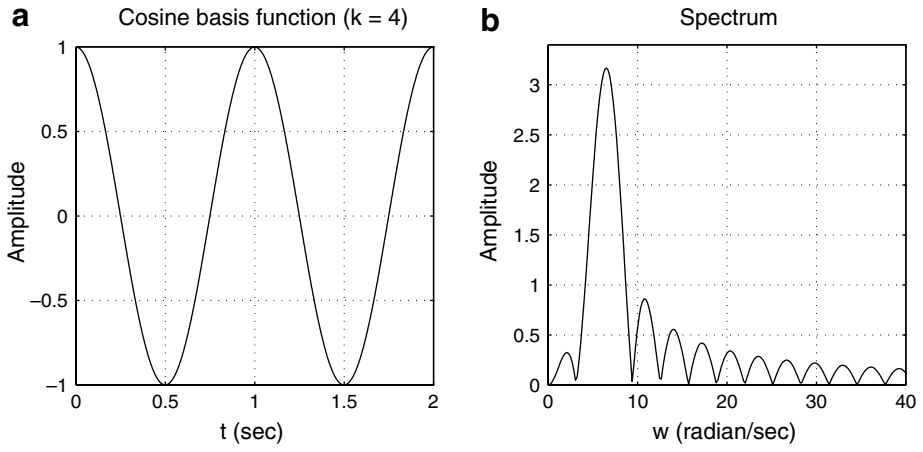

Fig. 6. Spreads of the basis functions of Cosine-I Transformation: (a) in the time domain; (b) in the frequency domain.

for the MFDWCs and μ^l is the mean in the mel-scaled filterbank log-energy domain. Absence of a superscript indicates the mel-scaled filterbank energy domain, e.g., μ represents the mean for this case. The symbols $\sim$ and $\hat{}$ indicate noise and estimated noisy speech parameters, respectively. All variables in bold are matrices or vectors, and subscripts indicate elements of the vectors or matrices. T is the Discrete Cosine Transformation matrix for calculating the MFCCs or the Discrete Wavelet Transformation Matrix for the MFDWCs. T^{-1} is inverse transformation matrix for the DCT or

DWT. X^{T} represents transpose of X where X may be a matrix or vector. The PMC technique can be summarized for HMMs with multiple mixtures, diagonal covariance matrices, static and delta coefficients as follows.

Let M_{s} and M_{n} be the numbers of mixtures in a state for the clean speech and noise models, respectively. Then there will be $M_{\mathrm{s}} * M_{\mathrm{n}}$ mixtures in each noisy speech state. Let $w_{\mathrm{s}i}$, $w_{\mathrm{n}j}$, w_k be the ith, jth and kth mixture weights of clean speech, noise and noisy speech states, respectively, where $k = (M_{\mathrm{s}} - 1) * i + j$.

Then the kth mixture weight for the noisy speech is

$$w_k = w_{\mathrm{s}i} * w_{\mathrm{n}j}. \tag{10}$$

Noisy speech state parameters for each mixture can be calculated using a speech mixture and a noise mixture using the following steps:

(1) Inverse transformation to get to the log-energy domain:

$$\mu^{\mathrm{l}} = T^{-1}\mu^{\mathrm{c}}, \tag{11}$$

$$\Delta\mu^{\mathrm{l}} = T^{-1}\Delta\mu^{\mathrm{c}}, \tag{12}$$

$$\Sigma^{\mathrm{l}} = T^{-1}\Sigma^{\mathrm{c}}(T^{-1})^{\mathrm{T}}, \tag{13}$$

$$\Delta\Sigma^{\mathrm{l}} = T^{-1}\Delta\Sigma^{\mathrm{c}}(T^{-1})^{\mathrm{T}}. \tag{14}$$

(2) Exponential transformation:

$$\mu_i = \exp\left(\mu_i^{\mathrm{l}} + \Sigma_{ii}^{\mathrm{l}}/2\right), \tag{15}$$

$$\Delta\mu_i = \exp\left(\Delta\mu_i^{\mathrm{l}} + \Delta\Sigma_{ii}^{\mathrm{l}}/2\right), \tag{16}$$

$$\Sigma_{ij} = \mu_i\mu_j\left[\exp\left(\Sigma_{ij}^{\mathrm{l}}\right) - 1\right], \tag{17}$$

$$\Delta\Sigma_{ij} = \Delta\mu_i\Delta\mu_j\left[\exp\left(\Delta\Sigma_{ij}^{\mathrm{l}}\right) - 1\right]. \tag{18}$$

(3) Composition:

$$\hat{\mu} = \mu + g\tilde{\mu}, \tag{19}$$

where g is a gain matching term which is given by

$$g = \frac{E_{\mathrm{ns}} - E_{\mathrm{n}}}{E_{\mathrm{s}}}, \tag{20}$$

where E_{ns}, E_{n} and E_{s} are average noisy speech signal energy, average noise energy and average clean training speech energy, respectively.

$$\Delta\hat{\mu}_i = \gamma_i\Delta\mu_i + g\eta_i\Delta\tilde{\mu}_i, \tag{21}$$

$$\widehat{\Sigma} = \Sigma + g^2\tilde{\Sigma}, \tag{22}$$

$$\Delta\widehat{\Sigma}_{ij} = \gamma_i\gamma_j\Delta\Sigma_{ij} + g^2\eta_i\eta_j\Delta\tilde{\Sigma}_{ij}. \tag{23}$$

(4) Logarithm transformation:

$$\hat{\mu}_i^{\mathrm{l}} = \log(\hat{\mu}_i) - \frac{1}{2}\log\left(\frac{\widehat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1\right), \tag{24}$$

$$\Delta\hat{\mu}_i^{\mathrm{l}} = \log(\Delta\hat{\mu}_i) - \frac{1}{2}\log\left(\frac{\Delta\widehat{\Sigma}_{ii}}{\Delta\hat{\mu}_i^2} + 1\right), \tag{25}$$

$$\widehat{\Sigma}_{ij}^{\mathrm{l}} = \log\left(\frac{\widehat{\Sigma}_{ii}}{\hat{\mu}_i\hat{\mu}_j} + 1\right), \tag{26}$$

$$\Delta\widehat{\Sigma}_{ij}^{\mathrm{l}} = \log\left(\frac{\Delta\widehat{\Sigma}_{ii}}{\Delta\hat{\mu}_i\Delta\hat{\mu}_j} + 1\right). \tag{27}$$

(5) Linear transformation:

$$\hat{\mu}^{\mathrm{c}} = T\hat{\mu}^{\mathrm{l}}, \tag{28}$$

$$\Delta\hat{\mu}^{\mathrm{c}} = T\Delta\hat{\mu}^{\mathrm{l}}, \tag{29}$$

$$\Sigma^{\mathrm{c}} = T\widehat{\Sigma}^{\mathrm{l}}T^{\mathrm{T}}, \tag{30}$$

$$\Delta\widehat{\Sigma}^{\mathrm{c}} = T\Delta\widehat{\Sigma}^{\mathrm{l}}T^{\mathrm{T}}. \tag{31}$$

4. Weighting the contribution of coefficients for the total score based on the noise level of coefficients

Noise affects each coefficient differently. This is especially true for features that are local in the frequency domain such as MFDWCs. When one frequency band is corrupted by noise, only a few coefficients will be affected if the speech features represent local information as in MFDWCs. Even if the entire frequency band is corrupted, the noise level can be different for each frequency interval for real applications, so the various coefficients will be affected differently. Therefore, it may be necessary to weight the contribution of each coefficient to the total score based on the noise level of the coefficients.

When all assumptions are true for estimating the noise and noisy speech model parameters, there is no need to weight the contribution of each coefficient to the total recognition score based on SNR. However, the assumptions involved in estimating the noisy speech parameters prevent perfect estimation. Therefore, it may be better to weight the contribution of each coefficient based on the noise level.

Let $O^{\mathrm{c}}(t)$ be observation vector at time t. Then the state observation likelihood for the diagonal covariance case will be

$$\sum_{k=1}^{M} w_k \prod_{i=1}^{N}\left\{\frac{1}{\sqrt{2\pi}\hat{\sigma}_i^{\mathrm{c}}}\exp\left\{-\frac{1}{2}\left(\frac{O_i^{\mathrm{c}}(t) - \hat{\mu}_i^{\mathrm{c}}}{\hat{\sigma}_i^{\mathrm{c}}}\right)^2\right\}\right\}, \tag{32}$$

where w_k is the kth mixture weight, M is the number of mixtures, and N is the number of coefficients. Each mixture of the noisy speech states is estimated using a pair of mixtures, one of the noise state and one of a clean speech state. Therefore, the SNR value will be different for each mixture of the estimated noisy speech states. The weighting must thus be performed at the mixture level. One approach for weighting for a given mixture can be expressed as

$$\prod_{i=1}^{N}\left\{\frac{1}{\sqrt{2\pi}\hat{\sigma}_i^{c}}\exp\left\{-\frac{1}{2}\left(\frac{O_i^{c}(t)-\hat{\mu}_i^{c}}{\hat{\sigma}_i^{c}}\right)^2\right\}\right\}^{\alpha_i}, \tag{33}$$

where α_i is the weighting factor for the ith coefficient for a given mixture of a state. A smaller value of α_i means less weight for that coefficient. The value α_i must be chosen based on the SNR value of the coefficient. Instead of using Eq. (33), changing the variance $\left(\hat{\sigma}_i^{c}\right)^2$ based on the noise level will have a similar effect. In this case we increase the noisy speech variance $\left(\hat{\sigma}_i^{c}\right)^2$ in proportion to the variance of the noise in the cepstral domain. The variance $\left(\sigma_i^{c}\right)^2$ can be updated as

$$\left(\hat{\sigma}_i^{c}\right)^2 \leftarrow \left(\hat{\sigma}_i^{c}\right)^2 + \alpha\left(\hat{\tilde{\sigma}}_i^{c}\right)^2, \tag{34}$$

where α is a scalar and $\left(\hat{\tilde{\sigma}}_i^{c}\right)^2$ is the compensated noise variance (explained later) for the ith coefficient. There is no exact method for choosing the optimal a value. The choice of a depends on how well the noisy speech parameters are estimated. If every parameter is well estimated, 0 can be chosen for α. Choosing too large a value for α may degrade results.

It is more practical to update the covariance matrix in the log-energy domain. The covariance matrix can be updated in log-energy domain as

$$\hat{\Sigma}_{ij}^{l} \leftarrow \hat{\Sigma}_{ij}^{l} + \alpha\hat{\tilde{\Sigma}}_{ij}^{l} \tag{35}$$

and

$$\Delta\hat{\Sigma}_{ij}^{l} \leftarrow \Delta\hat{\Sigma}_{ij}^{l} + \alpha\Delta\hat{\tilde{\Sigma}}_{ij}^{l}, \tag{36}$$

where $\hat{\tilde{\Sigma}}_{ij}^{l}$ and $\Delta\hat{\tilde{\Sigma}}_{ij}^{l}$ are the compensated noise covariance coefficients. The difference between $\hat{\tilde{\Sigma}}_{ij}^{l}$ and $\tilde{\Sigma}_{ij}^{l}$ is that $\hat{\mu}_i$ is used instead of $\tilde{\mu}_i$ to calculate $\hat{\tilde{\Sigma}}_{ij}^{l}$. Similarly, $\Delta\hat{\mu}_i$ is used instead of $\Delta\tilde{\mu}_i$ for calculating $\Delta\hat{\tilde{\Sigma}}_{ij}^{l}$. This is necessary because the contribution of the noise variance to the noisy speech variance will be based on $\hat{\mu}_i$ and $\Delta\hat{\mu}_i$ not $\tilde{\mu}_i$ and $\Delta\tilde{\mu}_i$.

5. Experimental setup and results

As in the earlier PMC work by Gales and Young, we used the NOISEX-92 (Varga et al., 1992) database to evaluate and compare the performance of MFDWCs with MFCCs. It consists of separate training and test utterances in a single-speaker setup. The training and test files include 10 digits spoken 10 times each with approximately 1 s gap between them. Background noises include speech, STITEL, Lynx, F16, Car, Factory, and Operations Room recordings. The same test and training files were used for all experiments.

The speech signal was sampled at 16 kHz and analyzed with 32 ms hamming windows stepped by 10 ms. The FFT of each frame was used to calculate the power spectrum of the signal. For the computation of mel-scaled log filter-bank energies, 33 triangular mel-scaled band-pass filters were designed and implemented.

Our previous experimental results (Tufekci and Gowdy, 2000; Gowdy and Tufekci, 2000) using the TIMIT database have shown that symmetric wavelets give better results than antisymmetric wavelets, so we used symmetric wavelets for this work. The folded DWT was used to compute MFDWCs to decrease the high frequency artifact caused by discontinuities (Mallat, 1998) at border boundaries. All the wavelets used in this paper are shown in Figs. 3–5. MFDWC-1s to MFDWC-6s are computed using the filters associated with wavelet1 to wavelet6 (Cohen et al., 1992) shown in Figs. 3–5. The folded DWT for symmetric wavelets requires (Mallat, 1998) the input vector size to be $2^N + 1$, where N is an integer, and the output vector size to be $2^M + 1$, where M is an integer. In our case $2^N + 1 = 33$ and $2^M + 1 = 17$. Eight coefficients at scale four, four coefficients at scale eight, two coefficients at scale sixteen, and two coefficients at scale thirty two and the zeroth coefficient were used. The total number of static coefficients used was therefore seventeen. MFCCs were computed by taking the DCT of mel-scaled log filterbank energies. The first sixteen of the MFCCs as well as the zeroth coefficient were used to make the number of parameters equal for MFCCs and MFDWCs. All feature vectors also include delta coefficients.

Eight-state left-to-right (with no skip), single-mixture continuous-density HMM models with diagonal covariance matrices were constructed for each digit. The silence model is a one-state, five-mixture, continuous density HMM model. Therefore, each state of each digit model has five mixtures after utilizing the PMC technique. The same setup was also used for the matched system. Single mixture components were used for the matched

digit models of the single speaker, and five mixture components were used for the silence model. The HTK toolkit (Young et al., 1997) was used for training and testing. HMMs are first trained using the Viterbi algorithm. Then the Baum–Welch algorithm is used for fine tuning. The grammar consists of three silences followed by a digit in a loop. The silence model for each test condition was trained using the silence interval of the test files. The accuracy is calculated as $100 * [(N - S - D - I)/N]$ where N is the number of tokens. S, D, I are the number of substitution, deletion, and insertion errors, respectively. We conducted a series of experiments involving different noise types, different noise levels using MFCCs, and 6 different MFDWCs utilizing the PMC technique. Table 1 shows the results for 7 different noise types and 3 noise levels using MFCCs and MFDWC-2s (which gave the best results on average). The average recognition rates for each noise level was also included in Table 1. As seen from Table 1, the MFDWC-2s dramatically improved the performance for all noise types and for almost all noise levels. Improvement is especially significant for the −6 dB noise level. A comparison of the average recognition rates of the MFCCs and MFDWC-2s demonstrates significant overall improvement. As shown in Table 1, MFDWC-2s yielded 13.72 and 5.29 points improvement over MFCCs for −6 and 0 dB noise levels, respectively.

The recognition results for MFCCs and MFDWC-2s utilizing the weighting algorithm are given in Table 2. The value 0.2 was chosen as the weighting factor α for all experiments that utilize the weighting algorithm. After testing a range of values for α on the training data, the best average performance was obtained using a value of 0.2. It is likely that there is an optimal a value for each noise type and noise level. Overall, weighting is shown to improve performance. As shown in Table 2, weighting improved the performance for both MFCCs and MFDWC-2s on the average. For example the recognition rate increased from 80.43% to 83.29% for MFDWC-2s for −6 dB noise on the average, which represents a 14.6% error reduction. We also conducted experiments for the matched case (where training and testing conditions are the same). Table 3 shows the results for the matched case. As seen from Table 3 the matched system performed poorly compared to the PMC-based system for MFCCs and for MFDWC-2s.

Table 1
Recognition accuracies for the MFCCs and MFDWCs both utilizing the PMC technique for noisy speech

Noise type	MFCCs			**MFDWC-2s**		
	−6 dB	0 dB	+6 dB	**−6 dB**	**0 dB**	**+6 dB**
Speech noise	65	88	99	**83**	**97**	**100**
Lynx noise	57	87	100	**75**	**97**	**99**
STITEL noise	65	95	99	**83**	**100**	**100**
F16 noise	76	93	98	**87**	**95**	**99**
Factory noise	71	96	99	**78**	**98**	**100**
Car noise	75	94	99	**89**	**97**	**100**
O. Room noise	58	88	94	**68**	**94**	**99**
Average	66.71	91.57	98.29	**80.43**	**96.86**	**99.57**

Table 2
Recognition accuracies for the MFCCs and MFDWCs both utilizing the PMC technique and weighting algorithm for noisy speech

Noise type	MFCCs			**MFDWC-2 s**		
	−6 dB	0 dB	+6 dB	**−6 dB**	**0 dB**	**+6 dB**
Speech noise	65	90	100	**86**	**98**	**100**
Lynx noise	60	91	100	**79**	**99**	**100**
STITEL noise	70	97	100	**87**	**100**	**100**
F16 noise	73	94	98	**86**	**99**	**99**
Factory noise	67	89	98	**81**	**98**	**100**
Car noise	80	95	99	**93**	**98**	**100**
O. Room noise	56	94	96	**71**	**93**	**98**
Average	67.29	92.86	98.71	**83.29**	**97.86**	**99.57**

Table 3
Recognition accuracies for the matched system for the MFCCs and MFDWCs for noisy speech

Noise Type	MFCCs			**MFDWC-2s**		
	−6 dB	0 dB	+6 dB	**−6 dB**	**0 dB**	**+6 dB**
Speech noise	79	100	100	**76**	**100**	**100**
Lynx noise	76	98	99	**73**	**100**	**100**
STITEL noise	31	79	99	**43**	**78**	**99**
F16 noise	68	98	100	**73**	**96**	**100**
Factory noise	52	93	99	**52**	**93**	**97**
Car noise	88	98	100	**82**	**98**	**100**
O. Room noise	49	89	99	**52**	**91**	**99**
Average	63.29	93.57	99.43	**64.43**	**93.71**	**99.29**

These results suggest that the matched system is not the optimal system. There are two possible reasons that the matched system performed poorly. (1) The estimated noisy speech parameters are not reliable, since it is difficult to get a good estimate in extremely noisy cases. (2) The noise may be quasi-stationary. One stationary part of the noise and a particular digit may overlap in the training set, but they may not overlap in the test set, causing mistraining. However, there is no such a problem for estimating the noisy speech parameters using the PMC technique. We also conducted a series of experiments using different wavelets to find the wavelet that gives the best result and to examine if there is a relationship between performance and locality in the frequency and time domain. We know that MFDWCs represent local information in the frequency domain, but MFCCs do not. Therefore we expect that the noise affects the MFDWCs differently (some coefficients will be affected more and some coefficients will be affected less) but MFCCs will be affected at the same level since the full frequency band is used to extract the MFCCs. Therefore, we expect the MFDWCs perform better than MFCCs for noisy speech recognition both using the PMC technique.

From Figs. 3–5 it can be seen that some of the wavelets used to calculate the MFDWCs have better time–frequency resolutions than the other wavelets. However, there are not major time–frequency resolution differences between different wavelets compared to the time–frequency resolution differences between the wavelet basis functions and cosine basis functions.

The average recognition results are given in Table 4. All MFDWCs gave better performance than the MFCCs as expected. This result suggest that using local features (MFDWCs) improves the performance. The MFDWC-2s gave the best result on the average. It seems that the wavelet used to calculate the MFDWC-1s is more local than the others in the time and frequency domains as seen from the Fig. 3. However, the MFDWC-1s did not give the best results. The wavelet used to calculate the MFDWC-4s is less local than the others in the time and frequency domains as seen from Fig. 4, but there is no big difference between the performance of the MFDWCs-4 and the others. These results suggest that being jointly local in the frequency and time domains is not too critical. However, having different time and frequency resolutions is important as seen from big performance differences between MFCCs and MFDWCs. Recall that MFCCs have approximately same time and frequency resolutions but MFDWCs have different time and frequency resolutions. It seems that the experimental results contradict with what was claimed (better localization in the time and frequency domains will result in better performance for noisy speech recognition). The significant performance differences between the MFDWCs and MFCCs experimentally demonstrates that the use of local features in the frequency domain improves the performance for the noisy speech recognition using the PMC technique. However there are not big performance differences between different MFDWCs. We know that there are big differences between time–frequency localization of MFDWCs and MFCC but small time–frequency localizations differences between different MFDWCs used in this paper. The experimental results suggests that small differences between the time–frequency localization property of the features (MFDWCs) do not change the performance significantly as we see from Table 4. There are not big performance differences between different MFDWCs.

Table 4
Average recognition accuracies for the MFCCs and MFDWCs all of them utilizing the PMC technique

Feature	Average accuracies		
	−6 dB	0 dB	+6 dB
MFCCs	66.71	91.57	98.29
MFDWC-1s	**76.85**	**96.43**	**99.14**
MFDWC-2s	**80.43**	**96.86**	**99.57**
MFDWC-3s	**76.43**	**96.3**	**99.57**
MFDWC-4s	**77.42**	**95.29**	**99.71**
MFDWC-5s	**78.14**	**96.86**	**99.43**
MFDWC-6s	**78.71**	**97.14**	**99.43**

Table 5
Average recognition accuracies for the MFCCs and MFDWCs that all of them utilizing the PMC technique and weighting algorithm

Feature	Average accuracies		
	−6 dB	0 dB	+6 dB
MFCC	67.29	92.86	98.71
MFDWC-1s	**79.43**	**97**	**99.43**
MFDWC-2s	**83.29**	**97.86**	**99.57**
MFDWC-3s	**80.57**	**97.57**	**99.57**
MFDWC-4s	**79.29**	**97.14**	**99.57**
MFDWC-5s	**81.43**	**97.86**	**99.57**
MFDWC-6s	**81.43**	**98.14**	**99.57**

The performances of the MFCCs and MFDWCs utilizing the weighting algorithm are given in Table 5. As seen from Tables 4 and 5, the weighting algorithm significantly improved the performance, on the average, for the MFCCs and MFDWCs. These results suggest that weighting the contribution of each coefficient to the recognition score based on the noise level improves the performance.

6. Conclusion

In this paper we described a speech recognition system based on the PMC architecture and using MFDWCs parameters. It was shown that MFDWCs are superior to MFCCs for speech recognition in noisy environments when they are used in conjunction with the PMC technique. It was also shown that weighting the contribution of each coefficient, based on the noise level corresponding to that coefficient, further improves performance. The weighting scheme is based upon changing the variances of the mixtures of the HMM states based on each noise level. We tried six different wavelets which have different time–frequency resolutions for extraction of MFDWCs. All of them significantly improved the performance in comparison to MFCCs. MFDWC-2s gave superior results among the six MFDWCs types. We have also shown that a matched system may not yield better results than the PMC-based system because of training issues for the matched system.

References

Allen, J.B., 1994. How do humans process and recognize speech. IEEE Trans. Speech Audio Process. 2 (4), 567–577.

Beattie, V.L., Young, S.J., 1991. Noisy speech recognition using hidden markov model state based filtering. In: Proc. ICASSP, pp. 917–920.

Beattie, V.L., Young, S.J., 1992. Hidden markov model state based cepstral noise compensation. In: Proc. ICSLP, pp. 519–522.

Berstein, A.D., Shallom, I.D., 1991. An hypothesized wiener filtering approach to noisy speech recognition. In: Proc. ICASSP, pp. 913–916.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral substraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Bourlard, H., Dupont, S., 1996. A new ASR approach based on independent processing and recombination of partial frequency bands. In: Proc. ICSLP, pp. 422–425.

Cerisara, C., Haton, J.P., Mari, J.F., Fohr, D., 1998. A recombination model for multi-band speech recognition. In: Proc. ICASSP, pp. 717–720.

Chengalvarayan, R., 1999. Hierarchical subband linear predictive cepstral (hslpc) features for HMM-based speech recognition. In: Proc. ICASSP, pp. 409–412.

Cohen, A., Daubechies, I., Feauveau, J., 1992. Biorthogonal bases of compactly supported wavelets. Comm. Pure Appl. Math. 45, 485–560.

Cung, H.M., Normandin, Y., 1992. Noise adaptation algorithms for robust speech recognition. In: Proc. ESCA Workshop Speech Processing Adverse Conditions, pp. 171–174.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366.

Fletcher, H., 1953. Speech and Hearing in Communication. Krieger, New York.

Gales, M.J.F., Young, S.J., March 1992. An improved approach to the hidden markov model decomposition of speech and noise. In: Proc. ICASSP, pp. 233–236.

Gales, M.J.F., Young, S.J., 1993a. Cepstral parameter compensation for HMM recognition in noise. Speech Comm. 12, 231–240.

Gales, M.J.F., Young, S.J., 1993b. HMM recognition in noise using parallel model combination. In: Proc. EUROSPEECH, pp. 837–840.

Gales, M.J.F., Young, S.J., 1995a. A fast and flexible implementation of parallel model combination. In: Proc. ICASSP, pp. 133–136.

Gales, M.J.F., Young, S.J., 1995b. Robust speech recognition in additive and convolutional noise using parallel model combination. Comput. Speech Language 9, 289–307.

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

Ghitza, O., 1986. Auditory nerve representation as a front-end for speech recognition in a noisy environment. Comput. Speech Language 1, 109–130.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Comm. 16, 261–291.

Gowdy, J., Tufekci, Z., 2000. Mel-scaled discrete wavelet coefficients for speech recognition. In: Proc. ICASSP, vol. 3, pp. 1351–13554.

Hermansky, H., Tibrewala, S., Pavel, M., 1996. Towards ASR on partially corrupted speech. In: Proc. ICSLP, pp. 462–465.

Klatt, D.H., 1979. A digital filterbank for spectral matching. In: Proc. ICASSP, pp. 573–576.

Lockwood, P., Boudy, J., 1991. Experiments with a non linear spectral sub-tractor (nss), hidden markov models and the projection, for robust speech recognition in cars. In: Proc. EUROSPEECH, pp. 79–82.

Mallat, S., 1998. A Wavelet Tour of Signal Processing. Academic Press, San Diego, CA.

Mansour, D., Juang, B.H., 1989. The short-time modified coherence representation and noisy speech recognition. IEEE Trans. Acoust. Speech Signal Process. 37, 795–804.

Mellor, B.A., Varga, A.P., 1992. Noise masking in the MFCC domain for the recognition of speech in background noise. In: Proc. IOA, vol. 14, pp. 503–510.

Mirghafori, N., Morgan, N., 1998. Transmission and transition in multi-band ASR. In: Proc. ICASSP, pp. 713–716.

Tomlinson, M.J., Russell, M.J., Moore, R.K., Buckland, A.P., Fawley, M.A., 1997. Modeling asynchrony in speech using

elementary single-signal decomposition. In: Proc. ICASSP, pp. 1247–1250.

Tufekci, Z., Gowdy, J., 2000. Feature extraction using discrete wavelet transform for speech recognition. In: Proc. SOUTHEASTCON, pp. 116–123.

Tufekci, Z., Gowdy, J., 2001. Subband feature extraction using lapped orthogonal transform for speech recognition. In: Proc. ICASSP, vol. 1, pp. 149–152.

Varga, A.P., Moore, R.K., 1990. Hidden markov model decomposition of speech and noise. In: Proc. ICASSP, pp. 845–848.

Varga, A.P., Steenekan, H.J.M., Tomlinson, M., Jones, D., 1992. The noisex-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit.

Vaseghi, S., Harte, N., Milner, B., 1997. Multi resolution phonetic/segmental features and models for HMM based speech recognition. In: Proc. ICASSP, pp. 1263–1266.

Vetterli, M., Kovacevic, J., 1995. Wavelets and Subband Coding. Prentice-Hall, Inc.

Yang, R., Haavisto, P., 1995. Noise compensation for speech recognition in car noise environments. In: Proc. ICASSP, pp. 433–436.

Yang, R., Haavisto, P., 1996. An improved noise compensation algorithm for speech recognition in noise. In: Proc. ICASSP, pp. 49–52.

Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1997. The HTK Book. Entropic Cambridge Research Laboratory Ltd., version 2.1.