Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 142–152
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition ☆

Umit H. Yapanel, John H.L. Hansen *

The Center for Robust Speech Systems, University of Texas at Dallas, Department of Electrical Engineering, EC33, P.O. Box 830688, Richardson, TX 75083-0688, USA

Received 11 August 2004; received in revised form 28 July 2007; accepted 29 July 2007

Abstract

Acoustic feature extraction from speech constitutes a fundamental component of automatic speech recognition (ASR) systems. In this paper, we propose a novel feature extraction algorithm, perceptual-MVDR (PMVDR), which computes cepstral coefficients from the speech signal. This new feature representation is shown to better model the speech spectrum compared to traditional feature extraction approaches. Experimental results for small (40-word digits) to medium (5k-word dictation) size vocabulary tasks show varying degree of consistent improvements across different experiments; however, the new front-end is most effective in noisy car environments. The PMVDR front-end uses the minimum variance distortionless response (MVDR) spectral estimator to represent the *upper* envelope of the speech signal. Unlike Mel frequency cepstral coefficients (MFCCs), the proposed front-end does not utilize a filterbank. The effectiveness of the PMVDR approach is demonstrated by comparing speech recognition accuracies with the traditional MFCC front-end and recently proposed PMCC front-end in both noise-free and real adverse environments. For speech recognition in noisy car environments, a 40-word vocabulary task, PMVDR front-end provides a 36% relative decrease in word error rate (WER) over the MFCC front-end. Under simulated speaker stress conditions, a 35-word vocabulary task, the PMVDR front-end yields a 27% relative decrease in the WER. For a noise-free dictation task, a 5k-word vocabulary task, again a relative 8% reduction in the WER is reported. Finally, a novel analysis technique is proposed to quantify noise robustness of an acoustic front-end. This analysis is conducted for the acoustic front-ends analyzed in the paper and results are presented.

Keywords: Acoustic feature extraction; Robust speech recognition; Noise-robustness analysis

1. Introduction

Current state-of-the-art speech recognizers perform well under controlled and clean laboratory conditions. However, the performance gap between automatic speech recognizers and human listeners in real world settings is significant (Huang et al., 2001). Much of the progress in recent years have occurred by exploiting more complex algorithms with the help of faster computing (Hunt, 1999). On the other hand, little progress has been reported in the development of *core speech processing algorithms*. One good example is the wide use of an acoustic front-end which was proposed more than *two decades ago*. Almost all current speech recognition, as well as speaker recognition systems, use Mel-frequency cepstral coefficients (MFFCs) as the acoustic front-end (Davis and Mermelstein, 1980). Many researchers would agree that there is still a significant potential in formulating an acoustic front-end signal that will successfully maintain information needed for efficient speech recognition, *especially in noise*, while eliminating irrelevant information (Hunt, 1999).

☆ This work was supported by the US Air Force under Contract F30602-03-1-0110, and RADC under Contract (FA8750-05-C-0029), and in part by DARPA under Grant No. N66001-8906.
* Corresponding author. Tel.: +1 972 883 2910; fax: +1 972 883 2710.
E-mail address: John.Hansen@utdallas.edu (J.H.L. Hansen).
URL: http://crss.utdallas.edu (J.H.L. Hansen).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.07.006

The most crucial information needed for ASR is *a representation of the vocal tract transfer function* (VTTF) (Huang et al., 2001); therefore, capturing the VTTF while eliminating other extraneous information (e.g. such as *speaker dependent characteristics, especially pitch harmonics*) is a key requirement for a good acoustic front-end (Hunt, 1999; Gu and Rose, 2001). The VTTF is mainly encoded in the *short-term spectral envelope* (Jelinek and Adoul, 1999) and extracting the short-term spectral envelope accurately and robustly is important for high performance ASR. Moreover, incorporating perceptual considerations, such as *Mel and Bark scales*, into the acoustic front-end leads to improved accuracy (Davis and Mermelstein, 1980; Hermansky, 1990).

MFCCs (Davis and Mermelstein, 1980) have proven to be one of the most effective features for ASR. They are computed by applying a Mel-scaled filterbank either to the *short-term FFT magnitude spectrum* or to the *short-term LPC-based spectrum* to obtain a perceptually meaningful *smoothed gross spectrum*. Both FFT and LPC-based spectra, however, have a limited ability to remove undesired harmonic structure, especially for high pitched speech (Jelinek and Adoul, 1999; Gu and Rose, 2001). FFT-based MFCCs have also been shown to be less effective for stressed speech recognition than LP-based MFCCs primarily due to the changes in excitation characteristics (Bou-Ghazale and Hansen, 2000). Moreover, MFCCs are quite fragile in noise, and additional compensation, such as *feature enhancement and model adaptation*, is needed for acceptable performance in realistic environments (Hansen et al., 2001b; Yapanel et al., 2002; CU-Move, 2004).

Direct upper envelope estimation algorithms using pitch-synchronous analysis and peak-picking techniques for computing the upper envelope have shown promise. However, they are both computationally expensive and prone to non-robust behavior in noisy conditions (Jelinek and Adoul, 1999; Gu and Rose, 2001).

Minimum variance distortionless response (MVDR) spectrum has been shown to be a superior way of modeling speech compared to linear prediction (LP), especially for medium and high-pitched speech (Murthi and Rao, 2000). Its potential for application in a robust front-end is also explored in (Dharanipragada and Rao, 2001; Yapanel and Dharanipragada, 2003).

This paper proposes a new acoustic front-end based on the MVDR spectrum estimation method. The front-end is algorithmically very similar to the PMCC front-end but differs in the incorporation of perceptual considerations. In the earlier approaches (Dharanipragada and Rao, 2001; Yapanel and Dharanipragada, 2003), the perceptual scales were integrated through the use of a non-linearly spaced filterbank, in the PMVDR front-end, on the other hand, this step is eliminated by directly warping the FFT power spectrum. As demonstrated through experimentation, the PMVDR front-end produces better results for clean and adverse environments due to (Yapanel and Hansen, 2003) its ability to accurately model the upper spectral envelope at the perceptually important harmonics.

The remainder of this paper is organized as follows. In Section 2, basics of the MVDR spectral modeling are explained. Section 3 summarizes two recently proposed front-ends based on the MVDR method. We introduce the PMVDR front-end in Section 4 and explain its implementation in detail. Experimental evaluation of this new front-end is discussed in Section 5. After taking computational issues into consideration in Section 6, we conclude the paper with a noise-robustness analysis in Section 7.

2. Minimum variance distortionless response (MVDR) spectrum

The MVDR technique is widely used in the beamforming literature; however, its application to speech modeling is quite recent (Murthi and Rao, 2000). The MVDR spectrum is a good way of performing all-pole modeling on the speech spectrum. Unlike the FFT analysis where fixed bandpass filters are used regardless of the characteristics of the incoming signal, MVDR obtains the power spectrum estimates by using *data-dependent* bandpass filters (Capon, 1969). The clever design of the bandpass filters is the key to the good spectral characteristics of the MVDR approach. The signal power at a frequency ω_l, is computed by filtering the signal with a special filter, $h_l(n)$. The power at the frequency ω_l, is determined by measuring the output power of the filter $h_l(n)$. The Mth order FIR filter $h_l(n)$, is designed to minimize its output power subject to the constraint that its frequency response at the frequency of interest ω_l, have unity gain (Haykin, 1991; Murthi and Rao, 2000).

Obtaining the MVDR spectrum at all frequencies of interest may seem a rather costly operation because it requires a special filter design for each frequency. However, the MVDR spectrum for all frequencies can be conveniently represented in a parametric form and the parameters, $\mu(k)$, hence the MVDR spectrum, can be easily obtained by a modest non-iterative computation proposed by Musicus (1985). $\mu(k)$s are computed from the LP coefficients, $a_{\{i\}}$ and the prediction error variance P_e as follows:

$$P_{MV}^{(M)}(\omega) = \frac{1}{\sum_{k=-M}^{M}\mu(k)\mathrm{e}^{-\mathrm{j}\omega k}} = \frac{1}{|B(\mathrm{e}^{\mathrm{j}}\omega)|^2}, \tag{1}$$

$$\mu(k) = \begin{cases} \frac{1}{P_e}\sum_{i=0}^{M-k}(M+1-k-2i)a_i a_{i+k}^*, & k:0,\ldots,M, \\ \mu^*(-k), & k:-M,\ldots,-1. \end{cases} \tag{2}$$

2.1. Use of MVDR for all-pole voiced speech modeling

The word "modeling" in the speech recognition context refers to the ability of extracting reliable spectral envelopes. We will cite two important results here from (Murthi and Rao, 2000) and interpret them in order to better

understand why the MVDR spectrum is a preferred method of envelope estimation for voiced speech.

1. The MVDR spectrum of order $M = (2L - 1)$ provides an envelope that exactly models the powers of a symmetric discrete line spectrum consisting of L lines, or harmonics (Murthi and Rao, 2000). Therefore, for voiced speech, *although not an ideal line spectrum*, the MVDR of order $M = (2L - 1)$ will accurately model L harmonic powers that are spaced at multiples of the fundamental frequency.
2. The MVDR spectrum accurately models the peaks in the speech spectrum by successfully *connecting* the spectral peaks to form the spectral envelope (Murthi and Rao, 2000). Therefore, we can now rightfully refer to this envelope as the *upper spectral envelope* and claim that it will be robust in moderate-SNR noisy environments. This is due to the fact that most of the noise types that we encounter in the practical implementations (e.g. car noise) will have a more pronounced impact on the low energy portions of the spectrum, leaving the spectral peaks almost unaffected, and hence the MVDR envelope, should not be affected severely by additive noise.

For more details on the MVDR spectrum estimation and its suitability to speech modeling, we refer readers to (Murthi and Rao, 2000; Yapanel, 2005).

3. Previous MVDR-based acoustic front-ends

Several studies have considered incorporating the merits of MVDR spectrum into the speech recognition framework (Dharanipragada and Rao, 2001; Yapanel and Dharanipragada, 2003; Yapanel et al., 2003; Wolfel et al., 2003). The first use of MVDR in speech parameterization was for power spectrum estimation (Dharanipragada and Rao, 2001). In (Dharanipragada and Rao, 2001), the FFT spectrum in the MFCC computation method was simply replaced by a high-order MVDR spectrum. The remainder of the feature extraction algorithm was the same as the MFCC front-end, therefore these features are called MVDR-based MFCCs (MVDR-MFCCs) (i.e., see Fig. 1). Also Dharanipragada and Rao (2001) incorporated a second step for cepstral smoothing to reduce the variances of the feature vectors which was also shown to be useful. One obvious disadvantage is the high computational burden imposed by the high-order MVDR spectrum computation method and cepstral averaging. A generic diagram for computing the MVDR-MFCCs are given in Fig. 1.

A second study employing MVDR methodology for feature extraction is (Yapanel and Dharanipragada, 2003; Yapanel et al., 2003). The features developed in (Yapanel and Dharanipragada, 2003; Yapanel et al., 2003) are called *Perceptual MVDR-based cepstral coefficients* (PMCCs). In the PMCC front-end, the MVDR methodology is used

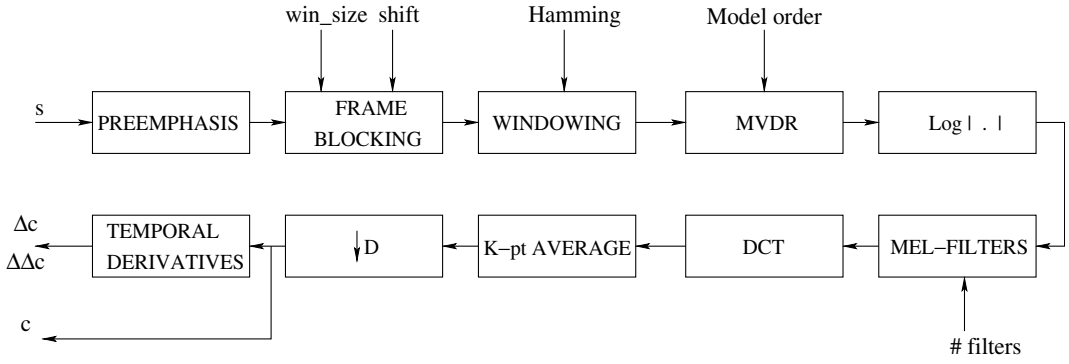

Fig. 1. Flow diagram of the MVDR–MFCC front-end.

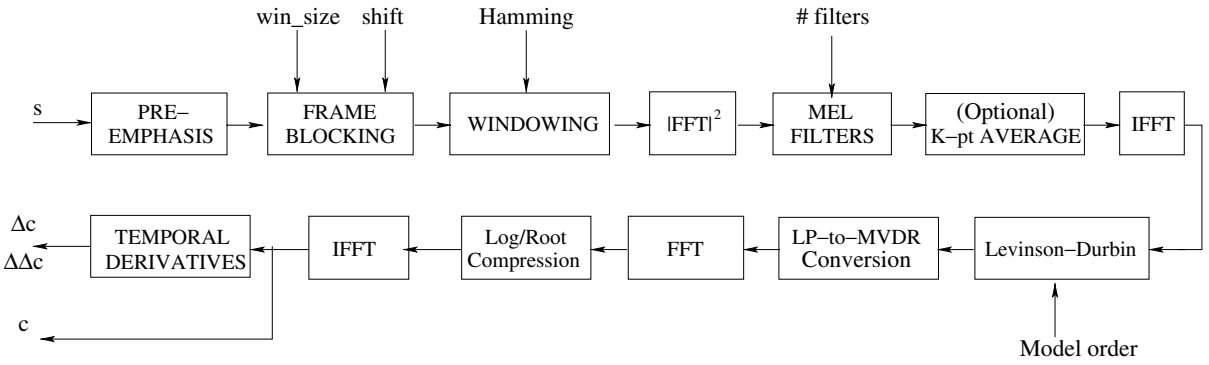

Fig. 2. Flow diagram of the PMCC front-end.

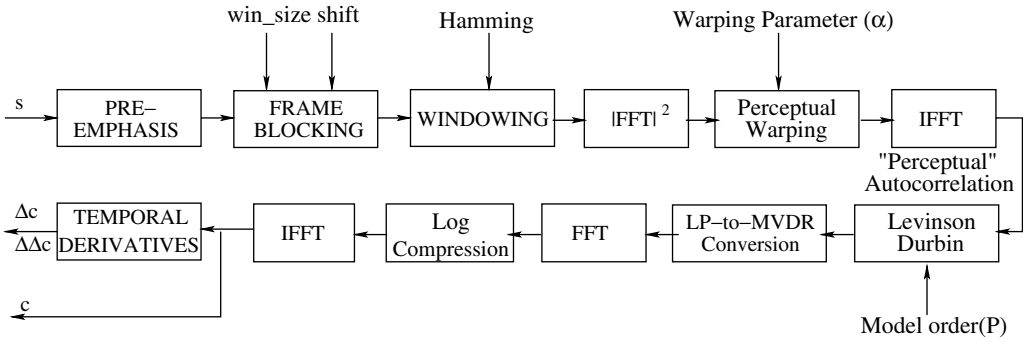

Fig. 3. Schematic diagram of PMVDR front-end.

for spectral envelope extraction rather than for spectrum estimation. It was shown that using the MVDR for spectral envelope extraction is more successful and yields better accuracy on the *IBM in-car speech recognition task* (Yapanel and Dharanipragada, 2003) and the *Wall Street Journal* (WSJ) task (Yapanel et al., 2003). The implementation of Perceptual MVDR-based Cepstral Coefficients (PMCCs) is very similar to PLP in that they both represent the spectral envelope using an all-pole model. However, the use of the MVDR-based and *not* the LP-based envelope in the all-pole modeling stage provides a measurable difference in performance for real car noise conditions (Yapanel and Dharanipragada, 2003). PMCCs were later utilized for clean speech recognition on the Wall Street Journal (WSJ) database and also shown to be more effective than the conventional MFCCs (Yapanel et al., 2003). Since it was shown that the PMCC approach substantially outperforms the MVDR–MFCC and PLP approaches (Yapanel and Dharanipragada, 2003; Yapanel et al., 2003), we consider only the PMCC front-end as baseline comparison in this study. A generic flow diagram for the PMCC front-end is given in Fig. 2.

4. Description of PMVDR

Previous approaches to integrating MVDR into speech parameterization for ASR involved using MVDR as a spectrum estimation (Dharanipragada and Rao, 2001) and as an envelope estimation technique (Yapanel and Dharanipragada, 2003; Yapanel et al., 2003). Different from the earlier approaches, PMVDR front-end completely removes the filterbank processing step and directly performs warping on the *FFT power spectrum*. The remainder of the algorithm is similar to the PMCC front-end (Yapanel and Dharanipragada, 2003; Yapanel et al., 2003). Our new approach is named PMVDR which stands for *perceptual MVDR cepstral coefficients*.

4.1. Direct warping of the FFT spectrum

Using a non-linearly spaced filterbank to incorporate perceptual traits into the acoustic front-end is a well-established technique (Davis and Mermelstein, 1980; Hermansky, 1990). The main aim of the filterbank is to average out the harmonic information (i.e., *the pitch*) that exists in the FFT spectrum and to track the spectral envelope. Since the filters are spaced closely at low frequencies, the effectiveness of filterbank in smoothing the pitch information is significantly reduced for high-pitch speakers. Therefore, the filterbank produces a gross spectrum that carries substantial pitch information which is not desirable for speaker-independent ASR applications (Gu and Rose, 2001). It was shown that MVDR is an appropriate spectral envelope modeling approach for a broad range of speech phoneme classes, especially for high-pitched speech (Murthi and Rao, 2000) Therefore, we can conclude that for an MVDR-based feature extraction algorithm, it is both useful and safe to remove the filterbank structure and incorporate the perceptual considerations directly into the FFT spectrum.

One way of incorporating perceptual considerations is to implement the perceptual scale through a *first order all-pass system* (Tokuda et al., 1994; Smith and Abel, 1999). This approach is simple and feasible for our purpose. In fact, both Mel and Bark scales are determined by changing the single parameter α of the system (Tokuda et al., 1994). The form, $H(z)$, and the phase response, $\hat{\omega}$, of the first order system are given as,

$$H(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1, \tag{3}$$

$$\hat{\omega} = \tan^{-1} \frac{(1 - \alpha^2)\sin(\omega)}{(1 + \alpha^2)\cos(\omega) - 2\alpha}, \tag{4}$$

where ω represents the linear frequency, while $\hat{\omega}$ represents the warped frequency. Here, the value of α controls the degree of warping. We are more interested in the non-linear phase response through which we implement the perceptual warping. For 16 kHz sampled signals, we set $\alpha = 0.42$ and 0.55 to approximate the Mel and Bark scales, respectively. For 8 kHz, these values are adjusted to $\alpha = 0.31$ and 0.42 (Tokuda et al., 1994). Bark scale performs more warping in the lower frequencies when compared to the Mel scale.

We note that alternative frequency spacing of the Mel filterbank has been shown to be effective for speech recognition under stress (Bou-Ghazale and Hansen, 2000) and automatic accent classification (Arslan and Hansen, 1996). Comparable shifts could also be incorporated into

the PMVDR computations for robust speech classification/recognition.

4.2. Implementation of direct warping

Warping via interpolation is a simple and fast method to implement direct warping. We would like to obtain the value of the power spectrum in the warped frequency space $\hat{\omega}$ by using its corresponding value in the linear frequency space, ω. The inverse relation that takes us from the warped to linear frequency space can be easily obtained from Eq (4) by *replacing* α with $-\alpha$,

$$\omega = \tan^{-1}\frac{(1-\alpha^2)\sin(\hat{\omega})}{(1+\alpha^2)\cos(\hat{\omega})+2\alpha}. \quad (5)$$

A step-by-step algorithm that describes how warping can be efficiently implemented via interpolation can be given as follows:

1. Take the FFT of the input speech frame of length N to obtain the FFT power spectrum. N should be selected as the nearest possible *power of 2*, thus providing N spectral points (i.e., $S[k], k = 0,\ldots,N-1$) in linear power spectrum space.
2. Calculate N linearly spaced spectral points over the warped frequency space by dividing the entire 2π warped frequency range into N equi-spaced points,
$$\hat{\omega}[i] = 2i\pi/N, \quad i = 0,\ldots,N-1. \quad (6)$$
3. Compute the linear frequencies and FFT indexes that corresponds to these warped frequencies using
$$\omega[i] = \tan^{-1}\frac{(1-\alpha^2)\sin(\hat{\omega}[i])}{(1+\alpha^2)\cos(\hat{\omega}[i])+2\alpha}, \quad i = 0,\ldots,N-1, \quad (7)$$
$$\hat{k}[i] = \frac{\omega[i]N}{2\pi}, \quad i = 0,\ldots,N-1. \quad (8)$$
4. For the final step, perform an interpolation of the nearest linear spectral values to obtain the warped spectral value
$$k_l[i] = \min(N-2, \hat{k}[i]), \quad i = 0,\ldots,N-1, \quad (9)$$
$$k_u[i] = \max(1, k_l[i]+1), \quad i = 0,\ldots,N-1, \quad (10)$$
$$\widehat{S}[i] = (k_u[i]-\hat{k}[i])S[k_l[i]] + (\hat{k}[i]-k_l[i])S[k_u[i]], \quad (11)$$
where $k_l[i]$ is the lower nearest linear FFT bin, $k_u[i]$ is the nearest upper linear FFT bin and $\widehat{S}[i]$ is the value of the warped power spectrum that corresponds to FFT bin i. Thus, the spectral value $\widehat{S}[i]$, at the warped frequency index $\hat{k}[i]$, is computed as the linear interpolation of nearest upper, $S[k_u[i]]$, and lower, $S[k_l[i]]$, spectral values in the linear frequency space.

4.3. PMVDR algorithm

The proposed PMVDR algorithm can be summarized as follows:

1. Obtain the perceptually warped FFT power spectrum via interpolation.
2. Compute the "perceptual autocorrelations lags" by taking the IFFT of the "perceptually warped" power spectrum.
3. Perform an Mth order LP analysis via Levinson–Durbin recursion using perceptual autocorrelation lags (Makhoul, 1975; El-Jaroudi and Makhoul, 1991; Haykin, 1991).
4. Calculate the Mth order MVDR spectrum using Eq. (2) from the LP coefficients (Murthi and Rao, 2000).
5. Obtain the final cepstrum coefficients using the straightforward FFT-based approach (Oppenheim and Schafer, 1989). In this approach, after obtaining the MVDR coefficients from the perceptually warped spectrum, we take the FFT of the parametrically expressible MVDR spectrum. After taking log, we apply IFFT to return back to the cepstral domain.
6. Take the first N, generally 12 excluding 0th cepstrum, cepstral coefficients as the output of the PMVDR front-end. This is the *cepstral truncation step*.

A flow diagram for the PMVDR algorithm is given in Fig. 3. The algorithm is integrated into our recognizer as the default acoustic front-end and the source code and executables can be obtained from CSLR web site (CSLR, 2004) together with SONIC (Pellom, 2001; Pellom and Hacioglu, 2003; CSLR, 2004).

4.4. Robust estimation of short-term spectral envelope

We ran an experiment to illustrate that the PMVDR envelope is in fact less susceptible to noise due to its upper envelope modeling property. For the same voiced sound frame, we computed MFCC and PMVDR cepstrum vectors for both clean and 5 dB car noise corrupted frames. We give clean (solid) and noisy (dash-dotted) cepstrum vectors in Fig. 4. Severe deviation of MFCCs from the clean case is apparent. The whole feature vector is moved upwards with the added car noise. This kind of deviation is rather *dangerous* because the deviations are substantial from the mean. The HMMs has both means and variances to model small variations in speech sounds and they are able to compensate for small deviations around the mean but they cannot tolerate this type of movement away from the mean. We computed a 73.9% average deviation from the mean in the noisy case from the clean case which quantifies the significance of the effect of noise on the MFCC cepstrum vector. In Fig. 4b, the variation of PMVDR cepstrum vector is given. The variations are small compared to the MFCC case and most importantly they are around the mean so the HMMs can cope with this type and amount of variation more easily. The average deviation from the mean in noisy conditions is 33.7% which is much less than the MFCCs' average variation. This experiment is a good evidence of the fact that upper envelope modeling property of the PMVDR front-end is indeed the key point to its

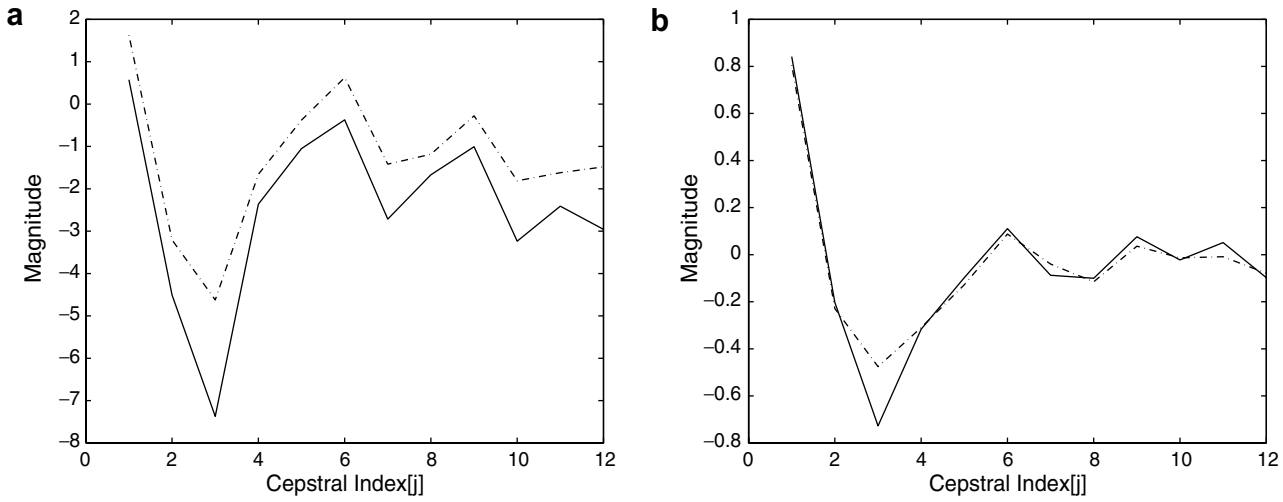

Fig. 4. Cepstrum values for a clean (solid) and 5 dB car-noise degraded (dash-dotted) voiced sound frame from /AA/ (a) variation of MFCCs, (b) variation of PMVDRs.

success for robust speech recognition. Note that in order to clearly illustrate the effects of the noise, neither *mean* nor *variance* normalization is utilized. However, throughout all the experimentation presented in this paper, we utilize cepstral mean normalization (CMN) by default for all front-ends.

5. Experimental framework

In order to test the effectiveness of the PMVDR front-end, recognition experiments were performed on *three different databases* that address different adverse conditions. The databases used in the simulations are: (a) *CU-Move Extended Digits Database* (CU-Move, 2004), for real noisy in-car environments; (b) *Speech Under Simulated and Actual Stress (SUSAS)* (LDC-SUSAS, 2004), for simulated stress conditions; and (c) *Wall Street Journal (WSJ)* (LDC-WSJ, 2004), for noise-free conditions. These databases cover a somewhat broad range of conditions which a recognizer might encounter in real-life applications.

5.1. General system description

For all experiments, we use SONIC (Pellom, 2001; Pellom and Hacioglu, 2003), the University of Colorado's Large Vocabulary Speech Recognition System. SONIC is a continuous density hidden Markov model (CDHMM) based recognizer. The acoustic models are decision-tree state-clustered HMMs with associated Gamma probability density functions to model state-durations. We used a window length of 25 ms and a skip rate of 10 ms by Hamming windowing the frame data before further processing. The 39 dimensional feature set contains 12 statics, deltas and delta-deltas along with normalized-log energy, delta and delta–delta energy. Cepstral Mean Normalization (CMN) was utilized on the final feature vectors for all front-ends considered in this paper. All HMMs have left-to-right topology with no skips and each state was represented by 6–24 mixtures depending on the available training data. During training, we fixed the state alignments for all front-ends, i.e. we did not re-align the training data with each front-end.

5.2. Experiments for noisy speech

For noisy speech experiments, we use the CU-Move Extended Digits Corpus which was collected in real car environments. The CU-Move project (CU-Move, 2004) aims to invent and develop car navigation systems that are reliable and employ a mixed-initiative dialog. This requires reliable speech recognition across changing acoustic conditions. There are 5 parts in the database; *command and control words*, *digit strings* being mostly phone numbers, *street addresses* with mostly spellings, *phonetically balanced sentences* and *Wizard of Oz* interactive navigation conversations. A total of 500 speakers produced over 600 GB of data during the 6-month collection effort across the United States. The database and noise conditions are analyzed in (Hansen et al., 2001a) in detail. We would like to emphasize that the noise conditions are changing with time in terms of SNR, stationarity and spectral structure. The SNR analysis presented in (Yapanel et al., 2002) revealed that the segmental SNR may change as much as 10 dB in real car environments for the only digits portion of this speech corpus.

A total of 60 speakers balanced across gender and age (18–70 yr old) were used in the training set. The test set contained another 50 speakers, again gender- and age-balanced. The HMMs were trained using SONIC's decision-tree HMM trainer resulting in 444 models with 513 total clustered states and around 10 K total Gaussians. The vocabulary size was 40. The dictionary is very convenient for telephone dialing applications since it contains many

Table 1
WERs (%) for CU-Move task with different front-ends

Gender/Systems	MFCC	PMCC	PMVDR	Rel. Imp.
Female	9.16	7.85	5.47	40.3
Male	13.22	12.03	10.16	23.1
Overall	11.12	9.87	7.74	30.4

necessary words like "dash", "pound", "sign" in addition to numbers. We compare the performance of PMVDR with previously proposed PMCC and the classical MFCC front-ends. As the first attempt we used a logical set of settings, i.e. Mel scale for the perceptual warp and a high enough LP order for LP analysis, for PMVDR. For the perceptual warp, we chose the *Mel warp* and pick the analysis order to be high enough to retain sufficient information for recognition while limiting speaker dependency of the features. The recognition performance for different front-ends in terms of WERs are given in Table 1 together with the relative improvements of PMVDR over the MFCC front-end. PMVDR decreases the WER by 30.4% relative to MFCC and 21.5% relative to PMCC front-ends.

We form a 17-speaker *development set* in order to optimize the PMVDR parameters. The development set does not overlap with train and test sets. The results verify that the PMVDR is more effective than MFCC and PMCC, especially for female speakers which are known to have high-pitched speech. We point out that while PMCC and PMVDR both use the MVDR representation for spectral envelope estimation, they differ in that PMVDR features use direct perceptual warping of the spectrum and does not require the use of a filterbank within the front-end. Hence the results support that using MVDR as an envelope estimation without a resolution decreasing filterbank leads to a better front-end for noisy environments.

5.2.1. Optimization of PMVDR parameters

PMVDR has *two parameters*, namely the LP order and *perceptual warp*, that can be optimized for different tasks. This could be an advantage for a front-end. Different tasks, such as accent classification (Arslan and Hansen, 1996) may require the use some other non-linear scaling of the spectrum for better performance. This scale optimization can be easily achieved within the PMVDR framework by simply adjusting the perceptual warp parameter (α). In the MFCC framework, one has to re-design the non-linearly spaced filterbank (Arslan and Hansen, 1996) to achieve the same effect. A second possibility is the integration of efficient speaker normalization algorithms within the PMVDR front-end. This possibility is demonstrated in (Yapanel and Hansen, 2005) where the perceptual warp factor is estimated separately for each speaker.

The two parameters of the PMVDR front-end can be logically related to two important characteristics of speech. The LP order used in the analysis is *directly related to the pitch period of the speaker*, for high-pitched speakers, we see less harmonics in the spectrum and we need a smaller order to sufficiently represent these harmonics (See (Murthi and Rao, 2000) for a detailed explanation of this claim). For low-pitched speakers, we see more harmonics in the spectrum and hence need a higher model order. This leads to a trade-off in model order selection since a compromise model order is needed to model both female and male speakers. A similar balance trade-off exists for the perceptual warp factor used to warp the FFT spectrum. The warp factor *can be related to the length of the vocal tract*. In fact, several speaker normalization algorithms exist using the same class of first-order warping functions for speaker normalization (McDonough et al., 1998). In the PMVDR case, these first-order warping functions are used to incorporate perceptual considerations. For females the vocal tract is *shorter* which, in turn, moves the formant frequencies higher. Therefore we do not want to expand low frequency region severely in order not to move the formants further away. However, for male speakers the situation is reversed. Male speakers have longer vocal tract lengths which causes the formants to move down in frequency and it is better to expand especially the low frequency space. *The optimal warp is again a mid-point which aligns average male formant with average female formant positions as closely as possible.*

We performed recognition experiments on the 17-speaker development set in order to determine the optimal settings for the CU-Move task. The variation of the WER with the LP analysis order is depicted in Fig. 5. During the experimentation we fixed the perceptual warp factor to be the Mel scale (i.e., $\alpha = 0.42$). WER for males and females follow a similar trend. For orders below 20, the WER increases as the LP analysis order is reduced and after 20, it is stabilized. This may seem contrary to our earlier

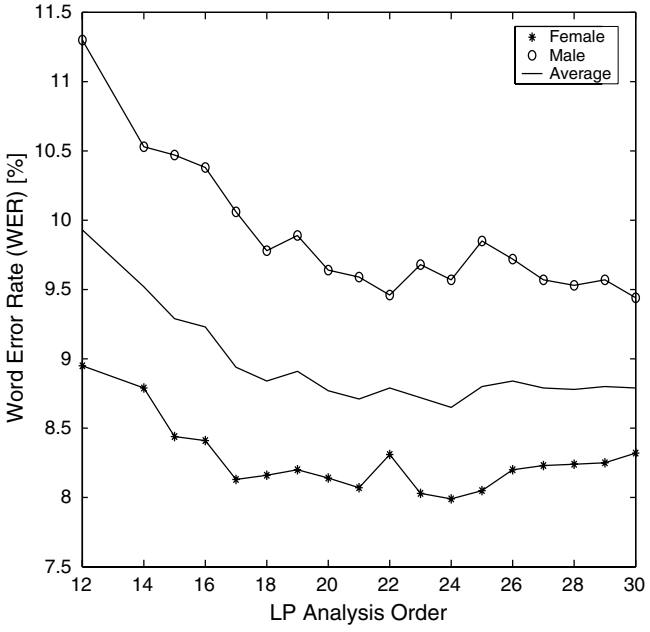

Fig. 5. Variation of WER (%) with LP analysis order, females (∗), males (○) and overall (solid).

discussion because of the final cepstral truncation part in the PMVDR computation. However, we are only using the first 12 PMVDR cepstrum coefficients regardless of the LP order by ignoring the rest. Increasing the LP order means increasing the detail in the spectrum. However, the cepstral truncation has a reverse smoothing effect on the spectrum. Therefore, after some order, added detail by the increased LP order will be smoothed out by the cepstral truncation stage and thus will have very limited effect on the WER. We see this behavior clearly in Fig. 5. After an order of 20, increasing the LP order does not affect the WER substantially. Orders lower than 20, however, do not adequately represent the vocal tract information needed for recognition. The optimal LP order is found to be 24 for this particular task, but as mentioned earlier any order higher than 20 works well for 16 kHz sampled speech. We note that increasing the LP order unnecessarily only wastes CPU resources with no explicit gain in the performance.

After fixing the LP order to be 24, we next perform experiments to optimize the perceptual warp factor. The variation of WER with perceptual warp factor is depicted in Fig. 6. The selected perceptual warp factor must balance the best performance between male and female vocal tracts. The Mel scale ($\alpha = 0.42$) is near optimal warp for female speakers but far from being optimal for male speakers. The optimal warp is $\alpha = 0.48$ for female speakers and $\alpha = 0.57$ for male speakers. Larger perceptual warp values for male speakers is an attempt to move the formants higher in frequency so that they are better aligned with those of female speakers. The optimal overall warp is again found to be $\alpha = 0.57$. At this warp, the WERs for female and male speakers are at the same level. Also, the car noise

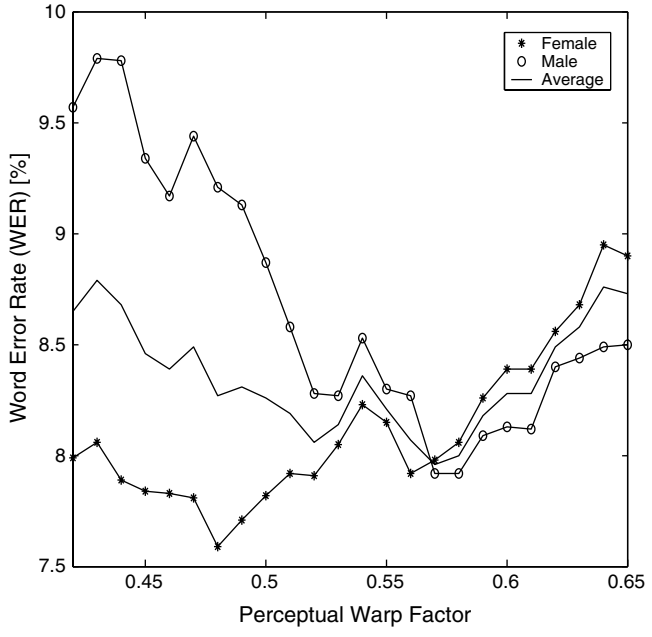

Fig. 6. Variation of WER (%) with the perceptual warp parameter (α), females (dashed), males (dash-dotted) and overall (solid).

Table 2
WERs (%) for CU-Move task with PMVDR Optimized settings

Gender/Systems	MFCC	PMCC	PMVDR	Rel. Imp.
Female	9.16	7.85	5.57	39.2
Male	13.22	12.03	8.76	33.7
Overall	11.12	9.87	7.11	36.1

as well as many other noise sources are concentrated at *lower frequencies* and this may have an impact on the optimal perceptual warp factor for the database under study (i.e., CU-Move corpus here).

Thus the optimal settings for the CU-Move task is found as $\alpha = 0.57$ and $M = 24$. The WER performance of the optimal settings on the test set are given in Table 2 together with the relative improvements over MFCCs.

5.3. Experiments for stressed speech

The performance of speech recognition systems degrade under the presence of stress (Hansen, 1996). Different speaking styles, such as fast, slow, question, soft, etc., have also a negative effect on the ASR performance. Therefore, it is informative to evaluate the proposed PMVDR front-end for speaker stress and different speaking styles as another adverse environment. Stressed speech in this context refers to the speech produced under environmental, emotional or workload stress. Depending on the type of stress, the *fundamental frequency, duration, intensity effects, glottal source*, and *vocal tract frequency structure* are all affected in different ways (Hansen, 1996). For example, for speech under angry conditions, the distribution of fundamental frequency expands substantially, the percentage of time spent in vowels and the corresponding amount of energy significantly increases at the expense of decrease in the percentage of time spent in consonants and consonant energy. The glottal spectral slope becomes more flat and formant locations as well as bandwidths are almost always statistically different from neutral conditions (Hansen, 1996; Bou-Ghazale and Hansen, 2000).

The speech data employed in this section is obtained from the SUSAS database (Hansen and Bou-Ghazale, 1997). SUSAS contains speech data produced under actual and simulated stress conditions across different speaking styles. Since the actual stressed speech part is also noisy, we decided to use only the simulated stress part in order to evaluate the robustness of the proposed PMVDR front-end. Simulated stress conditions include *neutral, angry, loud, clear, and, Lombard*. Lombard effect speech was obtained by having speakers listen to 85 dB SPL pink noise through headphones while speaking (i.e., recordings are noise-free). Different speaking styles include fast, slow, soft, and question. The simulated stress portion of the database consists of isolated words uttered by nine speakers. A common vocabulary set of 35 aircraft communication words make up over 95% of the database. These

words consist of highly confusable mono- and multi-syllabic words. Examples include /go-oh-no/, /wide-white/, and /six-fix/. Twelve tokens of each word in the vocabulary were spoken by nine native American speakers for the neutral conditions and two tokens for each stress and speaking style condition. Although the tokens are isolated words, we choose to train sub-word units in order to generalize our results to LVCSR applications. We trained left-to-right decision-tree state clustered HMMs using all available training data from all speakers under neutral conditions. Afterwards, the neutral-trained HMMs are tested against simulated stress conditions and different speaking styles. The HMM model set included 480 sub-word recognition units. The total number of Gaussians required was close to 5 K. We used the following simulated stress conditions and speaking styles in our evaluations; neutral, angry, high workload stress (cond 70), fast, slow, Lombard, loud, soft and question. We used MFCC front-end as the baseline. The PMCC acoustic front-end is also evaluated for this task with an LP order of 22. The PMVDR settings were $M = 24$ and $\alpha = 0.31$ (corresponds to the Mel scale) for the stressed-speech recognition task. We only ran two experiments using Mel and Bark scales for the perceptual warp factor and did not perform extensive optimization as in the CU-Move task. This was partly because of lack of data to form a separate development set. Note that using *the Mel scale for this task yielded slightly better results than the Bark scale*. The recognition performance and the relative improvements of PMVDR with respect to MFCC are summarized in Table 3. The Neutral condition represents the matched case in terms of stress.

The PMCC front-end yields a 10% improvement relative to the MFCC baseline. However, the PMVDR front-end is able to better address the stress and different speaking styles with a 27.8% relative improvement over the MFCC baseline. Although the amount of improvement depends on the stress type or speaking style considered, there are consistent improvements for every condition.

5.4. Experiments for clean speech

The task is noise-free read speech recognition with 5 K vocabulary on the well-known WSJ database. The sampling rate of the database is 16 kHz. The training set is the *SI-84* and the test set is the official *Nov'92 final eval set*. The final eval set includes three female and five male speakers with a total of 330 utterances. The total number of Gaussians was around 50 K for 612 decision-tree HMMs with around 2400 clustered states. The decoding was performed with *gender-independent* HMMs. We tabulated our results with MFCC, PMCC and PMVDR in Table 4 together with the relative improvements of PMVDR over MFCC. Note the improvement for *female* speakers clearly supporting the claim that MVDR is especially effective for medium and high-pitched speech (Murthi and Rao, 2000; Dharanipragada and Rao, 2001).

Table 3
WERs (%) for SUSAS database over different stress types

Type/Systems	MFCC	PMCC	PMVDR	Rel. Imp.
Neutral	3.66	3.97	3.02	17.5
Angry	50.79	50.79	36.98	27.2
Cond70	4.30	3.83	2.86	33.5
Fast	11.27	10.78	11.12	1.3
Slow	26.19	24.94	24.61	6.0
Lombard	12.07	9.67	5.86	51.5
Loud	38.72	32.07	26.49	31.6
Soft	18.90	17.63	12.66	33.0
Question	20.31	13.97	10.78	46.9
Overall	20.69	18.62	14.93	27.8

Table 4
WERs (%) for the WSJ November'92 Evaluation test set

Gender/Systems	MFCC	PMCC	PMVDR	Rel. Imp.
Female	6.99	5.93	5.35	23.5
Male	4.17	4.35	4.50	−7.9
Overall	5.22	4.93	4.82	7.6

6. Computational aspects

Computational performance can be considered under two main categories; namely the number of operations (NOP) required to compute the feature vector per frame, and the total real-time factor (RTF) required for the recognition test. The first is closely related to the algorithm of the feature set. The latter is tied to the properties of the features, such as suppression ability of noise and speaker variabilities. We summarize the NOP[1] and the real-time factors (RTFs) for MFCC, PMCC, and PMVDR front-ends on the WSJ task in Table 5.

The performance gain observed consistently for PMVDR comes at a computational price. The number of operations with respect to MFCC is now *doubled*. However, the PMVDR makes up for this loss in the search stage of the recognition. Better envelope modeling properties and robustness to noise and speaker variations leads to more efficient pruning in the search. This, in turn, yields a CPU gain of 14% relative over the MFCC baseline. Thus, we conclude that the PMVDR is also computationally tractable and suitable for both off-line and real-time ASR applications.

7. Noise-robustness analysis

Obtaining acceptable recognition performance in noise is a desirable property of a feature extraction algorithm. However, for a real in-car noisy database such as CU-Move, identifying the sources of improvement is rather difficult. We believe that an analysis should be performed, in addition to citing the final recognition results. We now

[1] Based on a 25 ms (or 400-sample) window at 16 kHz and a 50% overlap between consecutive frames.

Table 5
Computational complexity and RTFs for different front-ends

Step/# Operations	MFCC	PMCC	PMVDR
Windowing	400	400	400
$\|FFT\|^2$, $N = 512$	$512 \times \log_2(512) + 512$	$512 \times \log_2(512) + 512$	$512 \times \log_2(512) + 512$
Perceptual Warping	N/A	N/A	4×257
IFFT, $N = 512$	N/A	N/A	$512 \times \log_2(512)$
Filterbank ($P = 24$)	2×257	2×257	N/A
MVDR ($M = 22$)	N/A	2×22^2	2×22^2
FFT, $N = 128$	N/A	$128 \times \log_2(128)$	$128 \times \log_2(128)$
log	Ignored	Ignored	Ignored
IDCT	24×13	N/A	N/A
IFFT, $N = 128$	N/A	$128 \times \log_2(128)$	$128 \times \log_2(128)$
TOTAL NOP	6346	8794	12,888
TOTAL RTF	2.16	1.90	1.87

present an analysis technique which aims to quantify the noise robustness of an acoustic front-end.[2] In order to perform the noise robustness analysis, we will use *Segmental SNR (SegSNR)* (NIST, 2004) versus word error rate (WER) (Yapanel et al., 2002). For the proposed method of evaluation, we can summarize the steps as follows:

1. Segment the test set using an aligner tool. The segmentation level is basically a speech-silence detection. We used SONIC's aligner tool (Pellom, 2001) to align the data and extract speech-silence segmentation from the phone alignments.
2. Use NIST's SegSNR utility (NIST, 2004) to compute SegSNR for each utterance. The SegSNR calculation utility produces a sufficiently accurate SNR estimate for our purpose.
3. Average the SegSNR for each speaker and generate a scatter plot of the SegSNR versus WER for the entire test set.

The resulting plot is a measure of dependency between the SegSNR and WER. We propose to use the *correlation coefficient*, q, to evaluate the degree of this dependency. For a truly noise robust feature extraction algorithm, the correlation of SegSNR and WER should be close to 0. *The smaller the correlation coefficient, the more robust the acoustic front-end is to the background noise*.

We performed this analysis for three different acoustic front-ends, namely MFCC, PMCC and PMVDR. The correlation coefficients are summarized in Table 6. There is a negative correlation between the SegSNR and WER, as expected because while the SegSNR increases (data becomes less noisy) we would expect the WER to decrease. From the table, we observe that the smallest absolute value of the correlation coefficient is observed for PMVDR. This observation leads to the conclusion that the most fragile modeling strategy in noise is MFCC, while PMVDR delivers the greatest level of noise robustness among the three.

Table 6
Correlation coefficients of SSNR and WER for the 3 front-ends analyzed

Measure/Systems	MFCC	PMCC	PMVDR
Correlation Coef.	−0.29	−0.25	−0.19

8. Conclusions

In this paper, we proposed a new acoustic front-end, PMVDR, for ASR Systems. The proposed PMVDR front-end performs better than the conventional MFCC front-end and previously proposed PMCC front-end for a number of tasks including WSJ clean speech dictation task. Although PMVDR front-end is computationally more demanding than the MFCC front-end, it compensates for this loss in the search. Better acoustic modeling properties of the PMVDR front-end leads to considerable gains in real-time factors required for recognition.

Another important issue is the optimal values for the two parameters of the PMVDR front-end, namely the LP analysis order (M) and the perceptual warp factor (α). It was found that the LP order does not have much influence on the recognition accuracy provided that it is chosen larger than 20. Choosing a too large order increases computational load with no additional improvement in the WER. Therefore, we decided that an LP order of 22 or 24 is ideal for the PMVDR front-end. For the perceptual warp factor (α), we found different optimal values for different tasks. However, one useful observation is that if the task is noisy choosing the α close to the Bark scale provides substantial improvements over the Mel scale but for clean speech recognition tasks, using Mel scale provides slightly better results than using the Bark scale. However, we want to note that if the warp factor (α) is chosen within the [Mel, Bark] scale range, results are substantially better than the MFCC front-end and specific value of α is not very crucial.

[2] This approach to quantifying noise robustness was first proposed in (Yapanel and Hansen, 2003) and also used by CSLR for NRL-SPINE evaluation studies (Hansen et al., 2001b).

Acknowledgments

We express our great appreciation to Satya Dharanipragada from IBM T.J. Watson Research Center for his helpful discussions during stages of this research. We thank Bryan Pellom, formerly of University of Colorado for helpful discussions on the SONIC recognizer.

References

Arslan, L.M., Hansen, J.H.L., 1996. Language accent classification in American English. Speech Comm. 18 (4), 353–367.

Bou-Ghazale, S.E., Hansen, J.H.L., 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Trans. Speech and Audio Processing 8, 429–442.

CSLR, 2004. http://cslr.colorado.edu.

CU-Move, 2004. http://cumove.colorado.edu (also at http://crss.utdallas.edu).

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

Dharanipragada, S., Rao, B.D., 2001. MVDR-based Feature Extraction for Robust Speech Recognition, IEEE ICASSP-01: Inter. Conf. Acoust. Speech, Sig. Proc., pp. 3009-12, Salt Lake City, Utah.

El-Jaroudi, A., Makhoul, J., 1991. Discrete all-pole modeling. IEEE Trans. Signal Process. 39, 411–423.

Gu, L, Rose, K., 2001. Split-band perceptual harmonic cepstral coefficients as acoustic features for speech recognition, ISCA Interspeech-01/EUROSPEECH-01, Aalbrg, Denmark, pp. 583–586.

Hansen, J.H.L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition, speech communication. Special Issue on Speech Under Stress 20 (2), 151–170.

Hansen, J.H.L., Angkititrakul, P., Plucienkowski, J., Gallant, S., Yapanel, U., Pellom, B., Ward, W., Cole, R., 2001a. CU-Move: Analysis & Corpus Development for Interactive In-vehicle Speech Systems, Interspeech-01/EUROSPEECH-01, Vol. 3, Aalborg, Denmark, pp. 2023–2026.

Hansen, J.H.L., Sarikaya, R., Yapanel, U., Pellom, B., 2001b. Robust Speech Recognition in Noise: An Evaluation using the SPINE Corpus, Interspeech-01/EUROSPEECH-01, Vol. 2, Aalborg, Denmark, pp. 905–908.

Hansen, J.H.L., Bou-Ghazale, S.E., 1997. Getting Started with SUSAS: A Speech Under Simulated and Actual Stress database, ISCA EUROSPEECH-95, Rhodes, Greece, pp. 1743–1746.

Haykin, S., 1991. Adaptive Filter Theory. Prentice-Hall, Englewood Cliffs, NJ.

Hermansky, H., 1990. Perceptual Linear Prediction PLP Analysis of Speech. J. Acoustic. Soc. Am. 87 (4), 1738–1752.

Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice-Hall PTR, Upper Saddle River, New Jersey.

Hunt, M.J., 1999. In: Spectral Signal Processing for ASR, Vol. 1. Keystone, Colorado, pp. 17–26.

Jelinek, M., Adoul, J.P., 1999. Frequency-domain Spectral Envelope Estimation for Low Rate Coding of Speech. IEEE ICASSP-99: Inter. Conf. Acoust. Speech, Sig. Proc.. Phoenix, Arizona, pp. 1818–1821.

LDC-SUSAS, 2004, http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp? catalogId = LDC99S78.

LDC-WSJ, 2004, http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId = LDC93S6A.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proceedings of the IEEE 63, 561–580.

McDonough, J., Byrne, W., Luo, X., 1998. Speaker Normalization with All-pass Transforms, ISCA ICSLP-98: Internat. Conf. Spoken Lang. Proc., Sydney, Australia.

Murthi, M.N., Rao, B.D., 2000. All-pole modeling of speech based on the minimum variance distortionless response spectrum. IEEE Trans. Acoustic Speech Signal Process. 8 (3), 221–239.

Musicus, B.R., 1985. Fast MLM power spectrum estimation from uniformly spaced correlations. IEEE Trans. Acoustics Speech Signal Process. 33, 133–135.

NIST SPHERE Software Package, 2004, www.nist.gov.

Oppenheim, A.V., Schafer, R.W., 1989. Discrete-time Signal Processing. Prentice-Hall, Englewood Cliffs, NJ.

Pellom, B., 2001. SONIC: The University of Colorado Continuous Speech Recognizer, TR-CSLR-2001-01, Boulder, Colorado.

Pellom, B., Hacioglu, K., 2003. Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task, IEEE ICASSP-03: Inter. Conf. Acoust. Speech, Sig. Proc., Hong Kong, pp. 4–7.

Smith, J.O., Abel, J.S., 1999. Bark and ERB bilinear transforms. IEEE Trans. Speech Audio Process. 7 (6), 697–708.

Tokuda, K., Masuko, T., Kobayashi, T., Imai, S., 1994. Mel-generalized Cepstral Analysis-A Unified Approach to Speech Spectral Estimation, ISCA ICSLP-94: Inter. Conf. Spoken Lang. Proc., Yokohama, Japan, pp. 1043–1046.

Wolfel, M., McDonough, J., Waibel, A. 2003. Minimum Variance Distortionless Response on a Warped Frequency Scale, ISCA Interspeech-03/EUROSPEECH-03, Geneva, Switzerland, pp. 1021–1024.

Yapanel, U., 2005. Acoustic Modeling and Speaker Normalization Strategies with Application to Robust In-Vehicle Speech Recognition and Dialect Classification, PhD Thesis, Robust Speech Processing Group – CSLR, University of Colorado at Boulder.

Yapanel, U., Zhang, X., Hansen, J.H.L., 2002. High Performance Digit Recognition in Real Car Environments, ISCA Interspeech-02/ICSLP-02, Denver, Colorado, pp. 793–796.

Yapanel, U.H., Hansen, J.H.L., 2003. A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition, ISCA Interspeech-03/EUROSPEECH-03, Geneva, Switzerland, pp. 1281–1284.

Yapanel, U.H., Hansen, J.H.L., 2005. Towards an Intelligent Acoustic Front-end for Automatic Speech Recognition: Built-in Speaker Normalization (BISN), IEEE ICASSP-05: Internat. Conf. Acoust. Speech, Sig. Proc., Philadelphia, USA.

Yapanel, U.H., Dharanipragada, S., 2003. Perceptual MVDR-Based Cepstral Coefficients (PMCCs) for Noise Robust Speech Recognition, IEEE ICASSP-03: Internat. Conf. Acoust. Speech, Sig. Proc., Hong Kong, pp. 644–647.

Yapanel, U.H., Dharanipragada, S., Hansen, J.H.L., 2003. Perceptual MVDR-Based Cepstral Coefficients (PMCCs) for High Accuracy Speech Recognition, ISCA Interspeech-03/EUROSPEECH-03, Geneva, Switzerland, pp. 1829–1832.