# Perceptual MVDR-based Cepstral Coefficients (PMCCs) for High Accuracy Speech Recognition

*Umit H. Yapanel*[1], *Satya Dharanipragada*[2], *and John H.L. Hansen*[1]

(1) Robust Speech Processing Group, Center for Spoken Language Research
University of Colorado at Boulder, Boulder, CO, 80309, USA
{yapanel, jhlh}@cslr.colorado.edu, WEB: http://cslr.colorado.edu
(2) Human Language Technologies, IBM TJ Watson Research Center
Yorktown Heights, NY, 10598, USA
dsatya@watson.ibm.com

## Abstract

This paper describes an accurate feature representation for continuous clean speech recognition. The main components of the technique involve performing a moderate order *Linear Predictive* (LP) analysis and computing the *Minimum Variance Distortionless Response* (MVDR) spectrum from these LP coefficients. This feature representation, PMCCs, was earlier shown to yield superior performance over MFCCs for different noise conditions with emphasis on car noise [1]. The performance improvement was then attributed to better spectrum and envelope modeling properties of the MVDR methodology. This study shows that the representation is also quite efficient for clean speech recognition. In fact, PMCCs are shown to be a more accurate envelope representation and reduce speaker variability. This, in turn, yields a 12.8% relative word error rate (WER) reduction on the coombination of Wall Street Journal (WSJ) Nov'92 dev/eval sets with respect to the MFCCs. Accurate envelope modeling and reduction in the speaker variability also lead to faster decoding, based on efficient pruning in the search stage. The total gain in the decoding speed is 22.4%, relative to the standard MFCC features. It is also shown that PMCCs are not very demanding in terms of computation when compared to MFCCs. Therefore, we conclude that PMCC feature extraction scheme is a better representation of clean speech as well as noisy speech than MFCC scheme.

## 1. Introduction

Capturing the vocal tract transfer function (VTTF) from the speech signal while eliminating other extraneous *speaker dependent* information, such as pitch harmonics, is a key requirement for accurate speech recognition [2, 3]. It is well known that the vocal tract transfer function is mainly encoded in the short-term spectral envelope [4]; therefore, extracting the short-term spectral envelope accurately and in a manner invariant to noise is crucial for both clean and noise-robust speech recognition. It is also widely accepted within the speech recognition community that incorporating perceptual considerations, such as Mel and Bark scales, in the feature extraction process leads to improved accuracy and robustness [5, 6].

Mel-Frequency cepstral coefficients (MFCCs) have proven to be an effective set of features for speech recognition. In this method, a Mel-scaled filterbank is applied to either the short-term FFT spectrum or an LP-based spectrum to obtain a perceptually meaningful *smoothed gross spectrum*. This representation, however, has a limited ability to remove undesired harmonic structure, especially for high-pitch speech [3]. Furthermore, it has been observed that for high-pitch voiced speech, the formant frequencies are biased towards pitch harmonics and their bandwidths are therefore mis-estimated [3, 4, 2]. MFCCs are also expected to carry a good deal of speaker dependent information. The most obvious evidence of this is the fact that the same feature representation is widely used in speaker recognition systems. This also shows that the so-called *smoothed gross spectrum* cannot smooth out sufficient speaker-dependent information. While this property can be an advantage for speaker-dependent recognition, it certainly makes the use of MFCCs for speaker-independent recognition quite inefficient. FFT vs. LP-based MFCCs have also shown differences in performance due to speaker variability under noise, stress, and emotion [7]. Despite these observations, MFCCs are still the most widely used speech feature in state-of-the-art speech recognition systems.

In LP-based techniques, the spectral envelope is modeled by an all-pole filter whose coefficients are estimated by minimizing the Mean-Squared Error (MSE) between the spectrum and the LP filter's frequency response. The assumption is that the speech signal can be adequately modeled by the filter when the input is a single pulse or white noise [8]. However, this assumption does not hold exactly for voiced speech when the excitation is quasi-periodic [4]. Moreover, the MSE minimization is for the *speech spectrum itself*, not for its envelope [4]. Therefore as the analysis order increases, especially for high-pitch speakers, the envelope obtained from LP analysis tends to follow the fine structure of the speech spectrum and is biased towards strong harmonics. This causes the final speech features to carry a good amount of speaker dependent information. Furthermore, LP analysis is known to be highly sensitive to noise. Therefore, a technique that is better able to suppress the speaker dependent information in the spectrum and is robust to variations due to noise and environmental perturbations is necessary.

Direct upper envelope estimation using pitch-synchronous and peak-picking techniques for computing the *upper envelope* have shown promise but are computationally expensive and prone to non-robust behavior in noisy conditions [3].

Noise robustness of PMCCs was examined in [1]. The improvement was attributed to the fact that MVDR is very accurate and emphasizes the upper spectral envelope. It is a logical fact that noise corrupts the spectral valleys while leaving the spectral peaks (i.e. high energy regions) nearly unchanged; therefore a representation that makes use of this fact by modeling the upper spectral envelope is expected to be robust to variations due to additive noise. In this paper, we examine the performance of PMCCs on clean speech. Often, many noise-robust features, such as auditory features, can provide better performance in noisy conditions while causing reduced accuracy in noise-free

conditions. However, this is not the case with PMCCs. They are able to produce excellent results with both clean and noisy speech due to (1) their ability to model the upper envelope accurately thus yielding a performance gain in noisy conditions, (2) their ability to better suppress speaker-dependent information.

The paper is organized as follows. In the next section, we describe the MVDR spectral envelope estimation in detail. Section 3 deals with the algorithm of computing PMCC features. In section 4, we show, through a modified Linear Discriminant Analysis (LDA) [9], that PMCCs are better able to suppress speaker-dependent information than MFCCs. Section 5 gives the results of the experimental evaluations on the WSJ database. After considering computational issues in Section 6, we conclude the paper in Section 7.

## 2. MVDR Spectral Envelope Estimation

In the MVDR spectrum estimation method, the signal power at a frequency, $\omega_l$, is determined by filtering the signal by a specifically designed FIR filter, $h(n)$, and measuring the power at its output. $h(n)$, is designed to minimize its output power subject to the constraint that its response at the frequency of interest, $\omega_l$, has unity gain. This *distortionless constraint* ensures that the filter, $h(n)$, will let the input signal components with frequency $\omega_l$ pass through undistorted, and the minimization of the output power ensures that the remaining frequency components in the signal are suppressed in an optimal manner. This synergistic constrained optimization is a key aspect of the MVDR method which allows it to provide a lower bias with a smaller filter length than the Periodogram method [10]. Also, unlike the Periodogram method, the power is computed using all the output samples of the band pass filter, which also gives a reduction in the variance [11, 12].

This twofold optimization reduces the bias in the spectral samples and lowers the arbitrary variability in the resulting spectrum. These properties result in accurate spectrum estimation with desired statistical modeling characteristics. Other advantages of using MVDR include: (1) the resulting spectrum follows the upper spectral envelope closely, (2) the estimate is smoother than the LP spectra with similar conditions (better suppressing the speaker dependent information), (3) the formants are broader and not biased towards strong harmonics hence their position is more accurately estimated. For the same model order, LP tends to model the fine detail in the spectrum while MVDR mostly corresponds to upper envelope and contains almost *no excitation information*. This justifies the use of MVDR as the spectral envelope estimation technique for noise robust and high accuracy speech recognition. For more computational details of MVDR spectrum, the reader is referred to [11, 12, 1].

The $Q^{th}$ order MVDR spectrum can be parametrically written as

$$P_{MV}(\omega) = \frac{1}{\sum_{k=-Q}^{Q} \mu(k) e^{-j\omega k}} = \frac{1}{|B(e^{j\omega})|^2}. \quad (1)$$

The parameters, $\mu(k)$, can be obtained from a modest non-iterative computation using the LP coefficients, $a_k$, and prediction error variance, $P_e$ [13, 10].

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

The $(Q \nparallel 1$ coefficients, $\mu(k)$, completely determine the MVDR spectrum, $P_{MV}(\omega)$. Note that a *linear taper* or *triangular window* is used in the definition of $\mu(k)$ and this causes the MVDR spectrum to be *smoother* in appearance than the LP-based spectrum [13]. This makes the MVDR envelope a better representative of VTTF since it smoothes out unnecessary excitation details.

## 3. Computational Details of the PMCCs

In [12], the FFT-spectrum was simply replaced with a high order MVDR spectrum in the standard MFCC computation procedure. The method was shown to be very effective, especially for high-pitch speech in noisy conditions. This approach however has several problems. High order MVDR spectrum estimation is computationally very expensive. Furthermore the large-lag autocorrelation estimates required for the high order LP analysis are less reliable since they are forced to be estimated from a small (typically 25ms window) data sample causing high variance in the feature vectors. This makes a second smoothing step necessary via a K-point cepstrum averaging [12]. In order to overcome these problems, we take an approach similar to the Perceptual Linear Prediction (PLP) [6] method. Our PMCC approach utilizes MVDR as a spectral envelope estimation *not* as a spectrum estimation technique and this makes a remarkable difference in both implementation and performance. The steps needed for computation of PMCCs are explained below in detail;

1. Obtain filterbank output energies, $e[j]$, using the *FFT power spectrum* and $P$ Mel-scaled triangular filters,
2. Append the $P-2$ filterbank outputs (excluding $e[0]$ and $e[P-1]$) to the current output to get a symmetric vector of length $2(P-1)$,
3. Perform an IDFT (this reduces to IDCT since the input vector to IDFT is real and symmetric) on the vector obtained above to get $Q+1$ "perceptual" autocorrelations, $R[n]$. This can be implemented via a matrix multiplication of cosines;

$$R[n] = \frac{1}{M} \sum_{k=0}^{M-1} e[k] cos(2\pi kn/M), \;\; n = 0, 1, \ldots, Q \quad (3)$$

   where $M (= 2\;\; P-1)$.
4. Perform a $Q^{th}$ order LP analysis via Levinson-Durbin recursion [8, 13].
5. Obtain the MVDR spectrum , i.e. $\mu(k)$s, from the LP coefficients using Eq. (2) [11],
6. Convert the MVDR spectrum to cepstrum coefficients via the standard FFT-based approach [14].

The length of the FFT in the last step must be carefully chosen to prevent aliasing. As a final step, an optional cepstral smoothing (K-pt Averaging) can be performed to reduce the variance of the final feature vectors [12]. We ignored this step for clean speech recognition in order not to reduce the resolution of the feature vectors. A schematic diagram of the PMCC front-end illustrating the above steps is given in Fig. 1.

Fig. 2 compares the spectral envelopes represented by the first 13 MFCCs and PMCCs with the Mel-warped FFT power spectrum. The MFCCs are computed on the FFT power spectrum, *not on the magnitude spectrum*, so that both envelopes can be compared on the same scale. The figure gives useful insight on the characteristics of the envelopes. Fig. 2 (A) is for an unvoiced and (B) is for a voiced speech frame. The MFCC envelope shows arbitrary variations whereas the PMCC envelope is much smoother, thereby reducing speaker characteristics in the envelope. We mentioned that the MFCC envelope

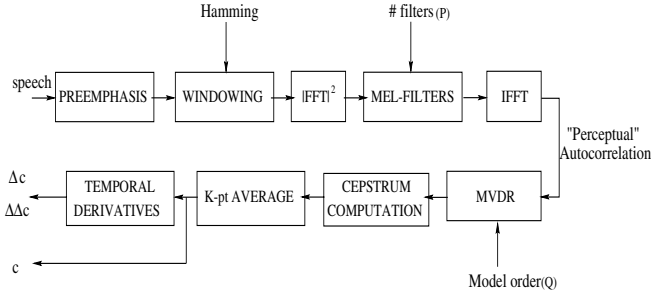


Figure 1: *Schematic diagram of the PMCC computation*

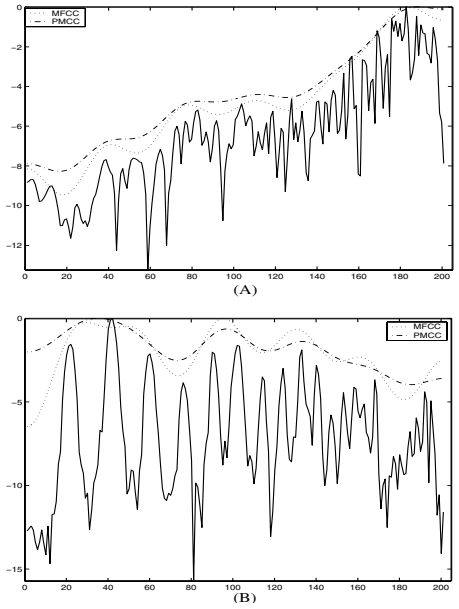


Figure 2: *Spectral envelopes for MFCCs (dotted) and PMCCs (dash-dotted) superimposed onto Mel-warped FFT power spectrum (solid) for (A) unvoiced, (B) voiced sounds of a female speaker from WSJ database* (x-axis denotes the *warped frequency index[k]* and y-axis denotes *normalized log-power* on both figures)

is biased towards strong harmonics and mis-estimates the formant bandwidths. Fig. 2 (B) is a good example. Consider the first formant, for the MFCC envelope *the formant is divided into two separate formants towards strong harmonics* whereas the PMCC envelope accurately represents only one formant at a more reasonable position. Similar arbitrary variations in the MFCC envelope falsely represents 5 formants while the PMCC envelope only shows 3 formants for the voiced frame. We conclude that, especially for voiced sounds, the PMCC envelope is more robust to strong harmonic structures and more accurately estimates the positions as well as the bandwidths of the formants.

## 4. A Measure of Inter-Speaker Variability

Our claim is that PMCCs are better able to suppress speaker dependent information than MFCCs. This section aims to evaluate the two feature extraction schemes in terms of their robustness to speaker variability. We use a modified Linear Discriminant Analysis (LDA) scheme proposed in [9] to evaluate the robustness to speaker variability. This scheme is basically a modified LDA in which we compute the within-class scatter matrix with respect to speaker variability; therefore, LDA objective function is now optimized with respect to speaker variations on phone classes.

Let $x^{(k)}$ be a D-dimensional feature vector output of feature extraction algorithm for discrete time $k$. We can define the mean of class $c$ for speaker $r$ as in Eq. (4).

$$m_{r,c} = \frac{1}{N_{r,c}} \sum_{k=1}^{N_{r,c}} x^{(k)}, \tag{4}$$

where the sum is over all $N_{r,c}$ feature vectors which belong to speaker $r$, $r \in \{1, \ldots, R\}$ and are assigned by a time alignment tool [16] to class $c$, $c \in \{1, \ldots, C\}$. Thus, this mean vector is for speaker $r$ and class $c$ and there is a total of $N = \sum_{c=1}^{C} N_c$ such mean vectors. $N_c$ denotes the total number of speakers that have spoken the phoneme $c$. For large enough databases, we can assume that every speaker have spoken every phoneme, thus $N_c = R$. Now, we apply Fisher Discriminant Analysis [17] and assume that each phoneme constitutes a class. We can then compute the class means by averaging over all the speakers as in Eq. (5).

$$m_c = \frac{1}{N_c} \sum_{r} m_{r,c}, \tag{5}$$

where the summation is over the speakers, $r$, who have spoken the phoneme $c$. We can then compute the global mean for the entire database via Eq. (6).

$$m = \frac{1}{N} \sum_{c=1}^{C} N_c m_c. \tag{6}$$

We compute between-class, $S_B$, and within-class, $S_W$, scatter matrices as in Eq. (7) and Eq. (8), respectively.

$$S_B = \frac{1}{N} \sum_{c=1}^{C} N_c (m_c - m)(m_c - m)^T. \tag{7}$$

$$S_W = \frac{1}{N} \sum_{c=1}^{C} \sum_{r} (m_{r,c} - m)(m_{r,c} - m)^T. \tag{8}$$

We would like to have feature vectors such that all vectors belonging to one class should be *compact* in the feature space *regardless of the speaker*. They should also be well-separated from the feature vectors of all other classes [9]. Two good measures of this property are the *determinant* and *trace* of $S_W^{-1} S_B$. The determinant is the product and the trace is the sum of the eigenvalues $\lambda_i$ of $S_W^{-1} S_B$ [9]. The interpretation for the determinant measure is straightforward. Since it is the evaluated value of the LDA objective function [17], we would like to maximize it. The interpretation for the trace measure is that the trace equals the sum of the variances in principal directions and is interpreted as the radius of the scattering volume. *The larger the trace is (i.e. the higher the class separability), the better separated the classes in the feature space*. This leads to the fact that [9] the higher the class separability, the lower the recognition error rate. The formula for computing determinant and trace measures are given below. Both measures are used to evaluate inter-speaker variability within phonemes in this study.

$$Dt = log(\prod_{i=1}^{C} \lambda_i). \tag{9}$$

$$Tr = \sum_{i=1}^{C} \lambda_i. \tag{10}$$

We give the evaluated values of the determinant and trace measures for MFCC and PMCC feature extraction schemes in Table 1. We conclude from the table that PMCCs show much less speaker variability, proving the claim that they better suppress speaker-dependent information than the MFCCs.

Table 1: *Det. and Trace measures for MFCCs and PMCCs.*

| Measure/Systems | MFCCs | PMCCs |
|---|---|---|
| Det. (Dt) | -74.57 | -70.44 |
| Trace (Tr) | 87.02 | 88.41 |

## 5. Experimental Evaluation

We use Sonic [16], the Univ. 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. The task is $5K$-vocabulary clean speech recognition on the WSJ database sampled at 16 $kHz$. The training set is the *SI-284* and the test sets are the official *WSJ0 5K Nov'92 dev* and *Nov'92 eval sets*. The *dev set* includes 4 female and 6 male speakers with a total of 410 utterances. The *eval set* includes 3 female and 5 male speakers with a total of 330 utterances. The 39 dimensional feature vector contains 12 statics, 12 deltas and 12 delta-deltas along with energy, delta and delta delta energy. We used a window length of $25ms$ and a skip rate of $10ms$ by Hamming windowing the frame data before further processing. 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. The total number of Gaussians was around $120K$ for 665 decision-tree, state clustered HMMs. We first trained gender-independent models and then adapted these models using available training data to male and female speakers. Decoding was performed via these adapted gender-dependent model sets [16]. We used $P$=33 filters in the Mel-filterbank and $Q$=24 as the model order for LP analysis in the computation of PMCCs which were optimized on an earlier task. We tabulated our results with MFCCs and PMCCs in Table 2 together with the relative improvements of PMCCs over MFCCs. The relative improvements are computed on the combination of both test sets. Note the outstanding improvement for *female* speakers. This result clearly supports the claim that MVDR is especially efficient for medium and high-pitch speech [11, 12, 1]. The results are also favorably comparable to those reported in [18].

Table 2: WERs(%) for WSJ **dev**/*eval* test sets.

| Gender/Systems | MFCCs | PMCCs | Avg. Rel. Imp. |
|---|---|---|---|
| Female | **3.9**/*4.5* | **3.1**/*3.9* | **14.6** |
| Male | **5.5**/*4.4* | **4.9**/*4.0* | **10.0** |
| **Overall** | **4.9**/*4.5* | **4.2**/*3.9* | **12.8** |

## 6. Computational Performance

Computational performance can be considered under two main categories; namely the number of operations required to compute the feature vector per frame, and the total time required for the recognition test. The first is closely related to the algorithm of the feature set. The latter, on the other hand, is tied to the properties of the features, such as suppression ability of noise and speaker variabilities. We summarize the number of operations (NOP)[1] and the real-time factor (RTF) for both MFCC and PMCC feature extraction schemes in Table 3.

Although, PMCCs require approximately 36% more NOP per frame than MFCCs, it compensates for this loss in the recognition stage. Better models, in the sense of suppressing speaker variability, lead to improved search and faster pruning, yielding a 22.4% faster decoding. This is also a remarkable improvement, especially for real- time systems.

Table 3: *NOP and RTF for WSJ task with MFCCs and PMCCs.*

| Comp./Systems | MFCCs | PMCCs | Rel. Imp.(%) |
|---|---|---|---|
| NOP | $\sim$11000 | $\sim$15000 | **-36** |
| RTF | 2.32 | 1.80 | **22.4** |

[1] Based on a 25ms (or 400-sample) window

## 7. Conclusions

In this paper, we extended the PMCC approach, which was earlier proposed for robust speech recognition, to the clean speech recognition. It is rarely the case that a feature representation be robust for both noisy and clean speech. The PMCC envelope is shown to be a more accurate representation than the MFCC envelope especially for highly harmonic voiced speech. A modified LDA analysis showed that PMCCs are also better able to suppress speaker dependent information yielding less speaker variability than the MFCCs. This, in turn, leads to more efficient search and pruning which results in a decoding speed gain of 22.4%. It is concluded that PMCCs are ideal for recognition of both clean and noisy speech. Thus, the PMCC features are strong candidates to replace MFCCs in future recognition systems.

## 8. Acknowledgments

U. H. Yapanel would like to thank B. Pellom of CSLR for his help with Sonic [16] and for very fruitful discussions.

## 9. References

[1] Yapanel, U. H. and Dharanipragada, S., "Perceptual MVDR-Based Cepstral Coefficients (PMCCs) for Noise Robust Speech Recognition", *Proc. ICASSP'03*

[2] Hunt, M. J., "Spectral Signal Processing for ASR", *Proc. ASRU'99*

[3] Gu, L. and Rose, K., "Perceptual Harmonic Cepstral Coefficients as the Front-end for Speech Recognition", *Proc. ICSLP'00*

[4] Jelinek, M. and Adoul, J. P., "Frequency-domain Spectral Envelope Estimation for Low Rate Coding of Speech"", *Proc. ICASSP'99*

[5] Davis, S. B. and Mermelstein, P., "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," *IEEE TASSP, Vol 28, 1980.*

[6] Hermansky, H. "Perceptual Linear Prediction (PLP) Analysis of Speech" *JASA, pp 1738-1752, 1990.*

[7] Bou-Ghazale, S. E., Hansen, J. H. L., "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress," *IEEE TSAP, vol. 8, pp. 429-442, July 2000.*

[8] Makhoul, J., "Linear Prediction: a Tutorial Review", *Proc. of IEEE, vol. 63, no.4, 1975*

[9] Haeb-Umbach, R., "Investigations on Inter-Speaker Variability in the Feature Space", *ICASSP'99*

[10] Marple, S. L., Jr., *Digital Spectral Analysis with Applications,* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[11] Murthi, M. N., and Rao, B. D., "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE TSAP, May 2000.*

[12] Dharanipragada, S. and Rao, B. D., "MVDR-based Feature Extraction for Robust Speech Recognition", *Proc. ICASSP'01*

[13] Haykin, S., *Adaptive Filter Theory,* Prentice-Hall, Englewood Cliffs,NJ, 1991.

[14] Oppenheim, A. V., Schafer, R. W., *"Discrete-time Signal Processing"* Prentice-Hall, Englewood Cliffs, NJ, 1989

[15] Sayed A. H., and Kailath, T., "A Survey of Spectral factorization methods," *Numerical Linear Algebra with Applications,* Vol. 8, pp. 467–496, 2001.

[16] Pellom, B., "Sonic: The University of Colorado Continuous Speech Recognizer", *Tech. Rep.TR-CSLR-2001-01*, CSLR, Univ. of Colo., March 2001.

[17] Duda, R. O., Hart, P., E., *"Pattern Classification and Scene Analysis",* John Wiley & Sons, NY, 1993

[18] Ney, H. and Welling, L. et al., "The RWTH Large Vocabulary Continuous Speech Recognition System" *Proc. ICASSP'98.*

[19] El-Jaroudi, A and Makhoul, J, "Discrete All-Pole Modeling," *IEEE Trans. Signal Processing,* Feb. 1991.

[20] Stoica P. and Moses R., *"Spectral Analysis"* Prentice-Hall, Englewood Cliffs, New Jersey, 1997.