PERCEPTUAL MVDR-BASED CEPSTRAL COEFFICIENTS (PMCCs) FOR ROBUST SPEECH RECOGNITION

Umit H. Yapanel[1]

Center for Spoken Language Research
University of Colorado at Boulder
Boulder, CO 80309-0594.
yapanel@cslr.colorado.edu

Satya Dharanipragada

Human Language Technologies
IBM T J Watson Research Center
Yorktown Heights, NY 10598.
dsatya@watson.ibm.com

ABSTRACT

This paper describes a robust feature extraction technique for continuous speech recognition. Central to the technique is the Minimum Variance Distortionless Response (MVDR) method of spectrum estimation. We incorporate perceptual information directly in to the spectrum estimation. This provides improved robustness and computational efficiency when compared with the previously proposed MVDR-MFCC technique [10]. On an in-car speech recognition task this method, which we refer to as PMCC, is 15% more accurate in WER and requires approximately a factor of 4 times less computation than the MVDR-MFCC technique. On the same task PMCC yields 20% relative improvement over MFCC and 11% relative improvement over PLP frontends. Similar improvements are observed on the Aurora 2 database.

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 [7, 9]. It is well known that the vocal tract transfer function is mainly encoded in the short-term spectral envelope [11]. Therefore, extracting the short-term spectral envelope accurately and in a manner invariant to noise is crucial for robust speech recognition. It is also widely accepted that incorporating perceptual considerations in the feature extraction process leads to improved accuracy and robustness [1, 4].

Mel-Frequency cepstral coefficients have proven to be an effective set of features for speech recognition. In this technique, a Mel-scaled filterbank is applied to the short-term FFT 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. Furthermore, it has been observed that for high pitch voiced speech, the formant frequencies are biased towards pitch harmonics and their bandwidths are therefore misestimated [9, 11, 7].

This work was performed as part of an IBM summer internship project

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 [11]. Moreover, the MSE minimization is for the speech spectrum itself not to its envelope, [11], 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. Furthermore, LP based spectra are known to be highly sensitive to noise.

Direct upper envelope estimation techniques using pitch-synchronous and peak-picking techniques for computing the *upper envelope* have shown a lot of promise but are computationally expensive and prone to non-robust behavior in noisy conditions [9].

In this paper, we describe a new feature extraction technique for continuous speech recognition. Central to the technique is the Minimum Variance Distortionless Response (MVDR) method of spectrum estimation. The method differs from previously proposed MVDR-MFCC technique in that perceptual considerations are incorporated directly in the spectrum estimation stage. This yields both accuracy and computational complexity improvements. MVDR models provide elegant envelope representations of the short-term spectrum of voiced speech. Furthermore, the MVDR spectrum is capable of modeling unvoiced speech, and mixed speech spectra. From a computational perspective, the MVDR modeling approach is also attractive because the MVDR spectrum can be simply obtained from a non-iterative computation involving the LP Coefficients, and can be based upon conventional time-domain correlation estimates.

2. MVDR SPECTRAL ENVELOPE ESTIMATION

In the MVDR spectrum estimation method, the signal power at a frequency ω_l is determined by filtering the signal by a

specially designed FIR filter $h(n)$ and measuring the power at its output. The FIR filter $h(n)$ is designed to minimize its output power subject to the constraint that its response at the frequency of interest, ω_l, has unity gain, namely

$$H(e^{j\omega_l}) = \sum_{k=0}^{M} h(k)e^{-j\omega_l k} = 1. \quad (1)$$

This constraint, known as the *distortionless constraint*, can be written as $\mathbf{v}^H(\omega_l)\mathbf{h} = 1$, where $\mathbf{h} = [h_0, h_1, \cdots, h_M]^T$, and $\mathbf{v}(\omega) = [1, e^{j\omega}, e^{j2\omega}, \cdots, e^{jM\omega}]^T$. Mathematically, the distortionless filter $h(n)$ is obtained by solving the following constrained optimization problem,

$$\min_{\mathbf{h}} \ \mathbf{h}^H\mathbf{R}_{M+1}\mathbf{h} \ \text{ subject to } \ \mathbf{v}^H(\omega_l)\mathbf{h} = 1. \quad (2)$$

where R_{M+1} is the $(M+1)\times(M+1)$ Toeplitz autocorrelation matrix of the data. The solution to this constrained optimization problem is [12, 6]

$$\mathbf{h_l} = \frac{\mathbf{R}_{M+1}^{-1}\mathbf{v}(\omega_l)}{\mathbf{v}^H(\omega_l)\mathbf{R}_{M+1}^{-1}\mathbf{v}(\omega_l)}. \quad (3)$$

The distortionless constraint ensures that the MVDR distortionless filter $h_l(n)$ will let the input signal components with frequency ω_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 that allows it to provide a lower bias with a smaller filter length than the Periodogram method. Also, unlike the Periodogram method, the power is computed using all the output samples of the bandpass filter, which gives a reduction in variance too.

One may think that designing a special FIR filter and computing its output power for all frequencies to obtain the MVDR spectrum is computationally too costly. Fortunately, there is a fast way of computing the MVDR spectrum. In fact, the MVDR spectrum for all frequencies can be conveniently represented in a parametric form as [12]

$$P_{MV}(\omega) = \frac{1}{\mathbf{v}^H(\omega)\mathbf{R}_{M+1}^{-1}\mathbf{v}(\omega)}. \quad (4)$$

Note that this represents the power obtained by averaging several samples at the output of the optimum constrained filter. This averaging results in reduced variance [5]. For computational purposes, the Mth order MVDR spectrum can be parametrically written as

$$P_{MV}(\omega) = \frac{1}{\sum_{k=-M}^{M}\mu(k)e^{-j\omega k}} = \frac{1}{|B(e^{j\omega})|^2}. \quad (5)$$

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 [12, 6]

$$\mu(k) = \begin{cases} \frac{1}{P_e}\sum_{i=0}^{M-k}(M+1-k-2i)a_i a_{i+k}^*, & k: 0,.,M \\ \mu^*(-k), & k: -M,.,-1 \end{cases} \quad (6)$$

The $(M+1)$ coefficients, $\mu(k)$, completely determine the MVDR spectrum $P_{MV}(\omega)$. From (5), the MVDR power spectrum can also be viewed as an all-pole model based power spectrum. Note also that a *linear taper* or *triangular window* is used in the definition of $\mu(k)$ and this causes MVDR spectrum be *smoother* in appearance than the LP-based spectrum [12]. This makes the MVDR envelope a better representative of VTTF since it smoothes out unnecessary excitation details. Empirical studies show that the MVDR method is indeed effective in removing pitch harmonics. Furthermore, a high order MVDR spectrum follows the upper envelope closely which is a desirable characteristic. Figure 1 compares the FFT-based, a low (15) and high order (60) LP-based and a 60th order MVDR-based spectral estimate for a typical voiced sound frame. We observe that the low order LP envelope is inaccurate whereas a high order LP envelope models the fine detail in the spectra. The MVDR envelope on the other hand accurately represents the upper envelope and contains almost no excitation information.

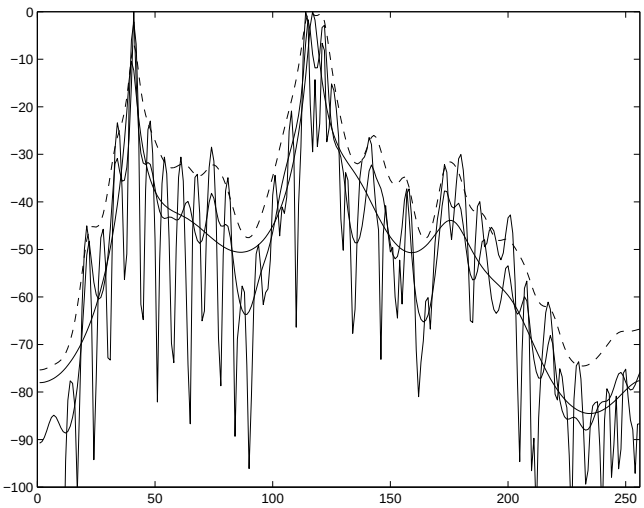

Figure 1 Spectral envelopes: LP (solid), MVDR (dashed)

2.1. Perceptual MVDR-based Cepstral Coefficients (PMCCs)

One approach to MVDR-based feature extraction would be to simply replace the FFT spectrum estimate with a high order MVDR spectrum estimate in the MFCC computation. This was investigated in [10] and was shown to be very effective owing to better spectrum estimates, especially for high pitch speech and noisy conditions. This approach however has several problems: high model order MVDR spectrum estimation is computationally very expensive. Furthermore the large-lag autocorrelation estimates required for the high order LPC analysis are less reliable since they are forced to be estimated from a small (typically 25ms window) data sample. This causes high variance in the feature vectors necessitating an additional smoothing step via temporal averaging. Finally, perceptual information was incorporated into the spectrum after the spectrum estimation, in the pre-

vious approach. It may be more advantageous to incorporate perceptual information in the spectrum estimation stage directly.

We address the above issues by adopting an approach similar to the Perceptual Linear Prediction technique [4], where perceptually modified autocorrelation estimates are obtained from a Mel-filtered spectrum. From these autocorrelation estimates we obtain the MVDR coefficients, $\mu(k)$s. This has two advantages. First the autocorrelation estimates are more reliable owing to the preceptual smoothing of the spectrum and thus the MVDR estimation is more robust. Second, the dimensionality of the MVDR estimation, and hence its complexity, is reduced owing the relatively smaller dimensionality of the Mel-Filterbank output.

The MVDR polynomial is a Laurent polynomial which is positive on the unit circle. Therefore,in order to compute the cepstral coefficients using the recursive procedure as in PLP we have to perform a Spectral Factorization of this polynomial, i.e. factor this polynomial as the modulus squared of a real algebraic polynomial. Several numerical techniques exist for Spectral Factorization[13]. Bauer method, Wilson method, Root calculation and Levinson-Durbin method are some common numerical methods. We experimented with the Levinson-Durbin method.

Cepstral coefficients can also be directly computed by computing the power spectrum from the MVDR polynomial using the FFT, taking the $\log$, and computing the inverse FFT or DCT. The length of the FFT must be carefully chosen to prevent aliasing. We observed that this method is both computationally efficient and is more accurate than the Levinson-Durbin method.

3. EXPERIMENTAL EVALUATION

We experimented with the PMCC technique on two databases – an automotive speech recognition application and the Aurora 2 database.

All experiments were conducted on a rank-based synchronous Viterbi decoder. The system uses context-dependent sub-phone classes which are identified by growing a decision tree using the training data and specifying the terminal nodes of the tree as the relevant instances of these classes. The raining feature vectors are poured down this tree and the vectors that collect at each leaf are modeled by a mixture of Gaussian pdf's, with diagonal covariance matrices. Each leaf of the decision tree is modeled by a 1-state Hidden Markov Model with a self loop and a forward transition. Output distributions on the state transitions are expressed in terms of the rank of the leaf instead of in terms of the feature vector and the mixture of Gaussian pdf's modeling the training data at the leaf. The rank of a leaf is obtained by computing the log-likelihood of the acoustic vector using the model at each leaf, and then ranking the leaves on the basis of their log-likelihoods.

For the in-car task, the training data consists of speech collected in a stationary and moving car at two different speeds – 30 mph and 60 mph. Data was recorded in several different cars with a microphone placed at a few different locations – rear-view mirror, visor and seat-belt. The training data was also appended by synthetically adding noise, collected in a car, to the stationary car data. Overall we have approximately 500 hours of training data.

The baseline system was trained using standard Periodogram-based MFCC vectors. Speech was coded into 25 ms frames, with a frame-shift of 15 ms. Since we are dealing with car noise, the 24 triangular Mel-filters were chosen in the frequency range [200Hz — 5500Hz]. Each frame was represented by a 39 component vector consisting of 13 MFCCs and their first and second time derivatives. Only the clean (stationary car) data was used to grow the decision tree. The Gaussian Mixture Models were trained on the entire data. Overall, we had 680 HMM states in our acoustic model. A total of just over 10,000 Gaussians model all the states.

Next, 39 dimensional PLP, MVDR and PMCC features with time-derivatives were generated. With these new feature streams, the means and the variances of the Gaussians and the transition probabilities of the HMM's were re-estimated using a Baum-Welch procedure to generate the corresponding acoustic models.

The test data comprises of 22 speakers recorded in a car moving at speeds 0 mph, 30 MPH and 60 mph respectively. Four tasks were considered: addresses, commands, digits and radio control. The test set, in total, has over 73,000 words in it.

Table 1. *WER for in-car data with different front-ends.*

Speed/Systems	MFCC	PLP	MVDR	PMCC
00mph	1.18%	1.14%	1.14%	1.13%
30mph	2.19%	1.93%	2.16%	1.97%
60mph	6.65%	5.93%	6.22%	4.92%
all	**3.34%**	**3.01%**	**3.18%**	**2.68%**

Table 2. *Relative improvement of PMCC with respect to MFCC, PLP and MVDR features.*

Speed/Systems	MFCC	PLP	MVDR
00mph	4.23%	0.87%	0.87%
30mph	10.04%	-2.07%	8.79%
60mph	26.01%	17.03%	20.90%
all	**19.76%**	**10.96%**	**15.72%**

PMCC improves the results approximately 20% with respect to the MFCC baseline, 11% with respect to the PLP and remarkably 15% with respect to the previous MVDR-MFCCs. This shows that the use of MVDR technique in PMCC is much more robust to noise than MVDR-MFCCs. Note that the improvement becomes much more apparent as the data becomes more and more noisy. In most cases, noise robustness results in some sacrifice in clean conditions, however, in the PMCC case there is a considerable amount of improvement in 00mph conditions which can be considered as being very close to clean conditions. This can

be attributed to the accurate envelope estimation achieved by the MVDR methodology.

For the Aurora 2 noisy digits database we use 156 subphone classes and a total of 3.5K Gaussians. As the training set we used multi-condition training and ran all the recognition tests on set A which consists of 4 different noise conditions (subway, babble, car and exhibition) at different SNR values. Averaging was done on 0dB-20dB SNR levels.

As seen in Table 3, PMCC provides significant improvements for almost all SNR levels. We also see a significant improvement in clean conditions. The averaged improvement for this artificially degraded database is 10.23%. It is worth noting that PMCC is much more effective for real data but it still gives considerable improvement over MFCC, PLP and MVDR-MFCC even for artificially degraded data.

Table 3. *WER results for Aurora 2 (Set A) and relative improvement of PMCC with respect to MFCC baseline.*

SNR	MFCC	PLP	MVDR	PMCC	Imp.
-5dB	65.60%	65.56%	65.28%	61.26%	6.62%
0dB	28.51%	28.20%	28.36%	25.23%	11.51%
5dB	9.27%	8.73%	9.65%	8.50%	8.31%
10dB	3.23%	3.28%	3.23%	3.26%	-0.93%
15dB	1.65%	1.86%	1.77%	1.50%	9.31%
20dB	1.29%	1.32%	1.22%	1.04%	19.38%
Clean	0.91%	0.78%	0.89%	0.78%	14.28%
0-20	**8.79%**	**8.68%**	**8.84%**	**7.89%**	**10.23%**

Table 4 provides a comparison of the computational complexity of the MFCC, MVDR-MFCC and the PMCC techniques in terms of the number of operations[1]. The PMCC technique can been to be far superior to the previously proposed MVDR-MFCC technique and requires only a modest increase in computations over the MFCC method.

Table 4. *Computational complexity of different front-ends*

System	# Operations	Increase(%)
MFCC	$\sim$6000	N/A
MVDR	$\sim$28000	370
PMCC	$\sim$8000	33

4. CONCLUSIONS

We described a new feature extraction technique, PMCC, for robust speech recognition. PMCC is based on MVDR envelope estimation technique which was shown to be accurate and robust. Results on two databases, an in car recognition task and the Aurora 2 database, showed that this technique provides significant improvements in accuracy and robustness. These improvements come with only a modest increase in computational complexity. The PMCC techniques is thus ideally suited for low resource speech recognition systems.

[1]based on a window size of 25ms at 11KHz sampling rate

5. ACKNOWLEDGEMENTS

The authors thank Prof. B.D. Rao of UCSD for many valuable discussions. U.H. Yapanel thanks his advisor John H.L. Hansen of CSLR for his continued support and encouragement for the summer project in which the work was performed.

6. REFERENCES

[1] S.B. Davis and P. Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing,* Vol 28, pp 357–366, 1980.

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

[3] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. on Speech and Audio Processing*, pp. 221–239, May 2000.

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

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

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

[7] M.J. Hunt "Spectral Signal Processing for ASR", *Proc. ASRU'99, December 1999*

[8] J. Makhoul "Linear Prediction: a Tutorial Review", *Proc. of IEEE, vol. 63, no.4, pp.561-580, 1975*

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

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

[11] M. Jelinek and J.P. Adoul "Frequency-domain Spectral Envelope Estimation for Low Rate Coding of Speech"", *Proc. ICASSP'99, pp.253-256*

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

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