A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition[§]

Umit H. Yapanel and John H.L. Hansen

Robust Speech Processing Group, Center for Spoken Language Research
Univ. of Colorado at Boulder, CO, 80309, USA
{yapanel, jhlh}@cslr.colorado.edu, WEB : http://cslr.colorado.edu

Abstract

The problem of reliable speech recognition for in-vehicle applications has recently emerged as a challenging research domain. This study focuses on the feature extraction stage of this problem. The approach is based on Minimum Variance Distortionless Response (MVDR) spectrum estimation. MVDR is used for robustly estimating the envelope of the speech signal and shown to be very accurate and relatively less sensitive to additive noise. The proposed feature estimation process removes the traditional Mel-scaled filterbank as a perceptually motivated frequency partitioning. Instead, we directly warp the FFT power spectrum of speech. The word error rate (WER) is shown to decrease by 27.3% with respect to the MFCCs and 18.8% with respect to recently proposed PMCCs on an extended digit recognition task in real car environments. The proposed feature estimation approach is called PMVDR and conclusively shown to be a better speech representation in real environments with emphasis on time-varying car noise.

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 [1, 2]. It is well known that the vocal tract transfer function is mainly encoded in the *short-term spectral envelope* [3]. Therefore, extracting the short term spectral envelope accurately and robustly (especially in additive noise) is crucial for robust speech recognition. It is also widely accepted within the speech recognition community that incorporating perceptual considerations, such as the Mel and Bark scales, into the feature extraction process leads to improved accuracy [4, 5].

Mel-Frequency cepstral coefficients (MFCCs) [4] have proven to be one of the most effective set of features for speech recognition. 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 the FFT and LPC-based spectrum, however, have a limited ability to remove undesired harmonic structure, especially for high pitch speech [2]. Furthermore, it has been observed that, for high pitch voiced speech, the formant frequencies are biased towards strong pitch harmonics and their bandwidths are therefore mis-estimated [2, 3, 1]. FFT-based MFCCs have also been shown to be less effective for stressed speech recognition than LP-based MFCCs [6]. Moreover, MFCCs are expected to carry a good deal of speaker dependent information. The evidence of this is that the same feature representation is commonly used in speaker recognition systems. It is also widely accepted that MFCC is quite fragile in noise and additional compensation such as feature enhancement and model adaptation is needed for acceptable performance in realistic environments [7, 8].

It is commonly agreed within the speech coding community that the spectral envelope, *not the gross spectrum*, represents the vocal tract transfer function [3, 2, 1]. For unvoiced sounds the spectral envelope and gross spectrum are similar. However, for voiced and transitional sounds, there can be a substantial mismatch [2]. Experimental evidence suggests that the *upper spectral envelope* which is obtained by sampling the FFT spectrum at the pitch harmonics is more effective, accurate, and reliable for speech recognition than the smoothed gross spectrum. The gross spectrum can be viewed as an averaging of the upper and lower spectral envelopes and thus more susceptible to noise and other environmental perturbations [2, 9]. Direct upper envelope estimation schemes using pitch-synchronous 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 [3, 2].

Minimum Variance Distortionless Response (MVDR) spectrum has been shown to be a superior way of modeling the speech compared to the linear prediction (LP), especially for medium and high-pitch speech [10]. It was earlier utilized for speech recognition in noise [11, 9]. In [11], the FFT spectrum, prior to the Mel-filterbank processing, was simply replaced by a high order (typically 80) MVDR spectrum. Although this approach yielded better results than MFCCs, it was rather computationally expensive. Another attempt to use the MVDR was made in [9]. The proposed PMCCs were similar to PLP [5] in terms of implementation and shown to improve recognition accuracy in real car noise conditions. The PMCC approach was able to use the upper envelope modeling property of the MVDR spectrum [10, 9] to some extent and this yielded substantial improvement.

The aim of the filterbank processing is to remove extraneous excitation information and to track the spectral envelope. However, this approach is shown to have a limited ability to remove strong harmonic structure for medium and high-pitch speech [1, 3, 2]. The MVDR methodology, on the other hand, can effectively model medium and high-pitch speech and tracks the upper envelope thereby excellently smoothing undesired excitation information. Therefore, it is feasible to completely remove the filterbank processing step from feature extraction process. This allows us to warp the FFT power spectrum directly, providing a better approximation to the perceptual scales. Thus, PMVDR methodology is able to produce superior results with noisy (and perhaps with clean) speech due to the two claims: its ability (1) to model upper envelope accurately yielding a performance gain in noisy conditions, (2) to better suppress speaker-dependent information yielding more accurate recognition and faster decoding in both clean and noisy conditions.

[§]This work was supported by DARPA through SPAWAR under Grant No.N66001-8906.

2. MVDR Spectral Envelope Estimation

For details of MVDR spectrum estimation and its previous uses for speech parameterization, we refer the reader to [10, 11, 9]. We only summarize the general properties and computational algorithm of the MVDR spectrum.

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. This 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 [12]. 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, 12]

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

3. PMVDR Formulation

Previous approaches to integrating the MVDR into speech parameterization for speech recognition involved using MVDR as a spectrum estimation [11] and as an envelope estimation technique [9]. It was shown that using MVDR methodology to estimate the spectral envelope leads to better performance [9]. In [9], perceptual considerations were integrated using the Mel-scaled filterbank at the expense of loosing some useful information. It is also a rough approximation to the perceptual scale since it samples the perceptual spectrum at the center frequencies of the filterbank. Furthermore, the filterbank is less effective in completely removing the harmonic excitation information from the spectrum. Because of these reasons and aforementioned benefits in Section 1, we remove the filterbank and perform warping directly on the *FFT power spectrum*. After obtaining a perceptual spectrum, the remainder of the estimation process is similar to the PMCC approach [9]. Our new approach is, however, named PMVDR which stands for *perceptual MVDR coefficients*.

3.1. Direct Warping of FFT Spectrum

It has been shown that implementing the perceptual scales through the use of a first order all-pass system is feasible [14, 15]. In fact, both Mel and Bark scales are determined by changing the only parameter, α, of the system [14]. The form, $H(z)$, and the phase response, $\beta(\omega)$, are given as;

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

$$\beta(\omega) = tan^{-1} \frac{(1-\alpha^2) sinw}{(1+\alpha^2) cosw - 2\alpha} \quad (4)$$

where ω represents the linear frequency while $\beta(\omega)$ represents the warped frequency. α controls the degree of warping. For 16 kHz sampled signals, $\alpha = 0.42$ and 0.55 approximate the Mel and Bark scales, respectively. For 8 kHz, these values are $\alpha = 0.31$ and 0.42 [14].

3.2. PMVDR Algorithm

Utilizing direct warping on the FFT power spectrum by removing the filterbank processing step leads to the preservation of almost *all* the information in the short-term speech spectrum. We can now summarize the remainder of the proposed PMVDR algorithm as follows;

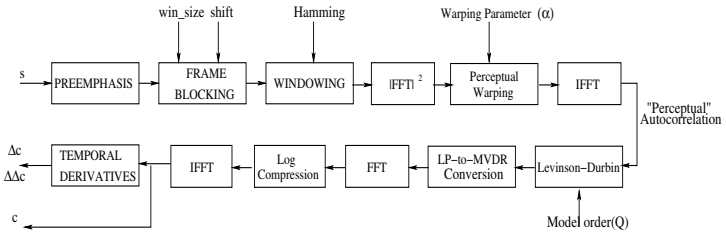

Figure 1: *Schematic diagram of PMVDR front-end computation*

1. Obtain the perceptually warped FFT power spectrum,
2. Compute the "perceptual autocorrelations" by utilizing the IFFT on the warped power spectrum,
3. Perform a Q^{th} order LP analysis via Levinson-Durbin recursion using perceptual autocorrelation lags [16, 13],
4. Calculate the Q^{th} order MVDR spectrum using Eq.(2) from the LP coefficients [10],
5. Obtain the final cepstrum coefficients using the straightforward FFT-based approach [17].

A flow diagram for the PMVDR algorithm is given in Fig. 1. The algorithm is integrated into our recognizer and the source code can be obtained from our web site together with Sonic [18, 19, 20].

A comparison of envelopes from different front-ends allows us to assess the trade-offs and merits of them. We show the envelopes respresented by the first 13 MFCCs, PMCCs and PMVDRs in Fig. 2. Here, MFCCs are computed from the power spectrum, not from the magnitude spectrum, for comparison purposes. Fig. 2 (A) compares the three envelopes for an unvoiced sound segment. Note the excellent match between the PMVDR envelope and warped spectrum. This can be attributed to the accurate perceptual warping achieved by the direct warping. PMCC *partly* corresponds to the upper envelope while PMVDR follows the upper envelope more closely. We also note that PMVDR is the smoothest envelope amongst the three. This reduces variances of the final model set thereby yielding sharper models. This, in turn, leads to *more accurate* and *faster* decoding. Fig. 2 (B) compares the envelopes for a voiced sound frame. MFCC and PMCC envelopes do not correctly match with the warped power spectrum because of the rough approximation by the filterbank, while PMVDR very accurately models the perceptual spectrum. We usually observe 2 or 3 formants in a voiced sound spectrum. Consider the envelope provided by MFCC. We hypothesize that it is *biased towards strong harmonics* and *this causes the formant bandwidths to be mis-estimated* by introducing spurious formants. We observe 4 formants in the MFCC envelope. PMCC envelope is a bit better suppressing one possibly false formant in the first broad formant structure. However, it still has *two formants divided towards the strong harmonics*. PMVDR excellently handles this very broad formant situation by showing two formants only at locations that match with the warped FFT power spectrum more accurately. It is this ability of PMVDR, in co-operation with the upper envelope property, that helps in high accuracy and noise-robust speech recognition. We would like to emphasize that the comments made here are based on our observations from real data. However, more thorough analyses should be performed on articial speech for which the formant positions and bandwidths are known to prove these conclusions.

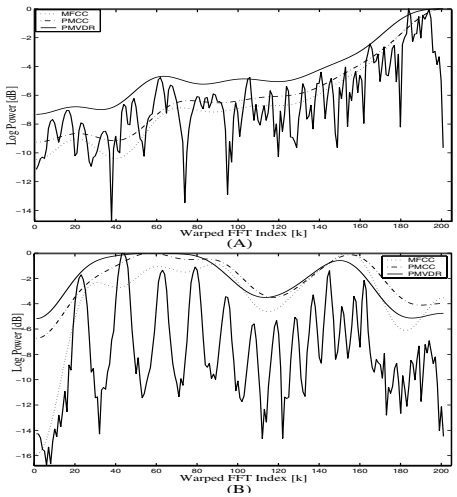

Figure 2: *Spectral envelopes for MFCC (dotted), PMCC (dash-dotted) and PMVDR (solid) superimposed onto Mel-warped FFT power spectrum for (A) unvoiced, (B) voiced sounds of a female speaker from WSJ database*

4. Experimental Evaluation

We use Sonic [18], the Univ. of Colorado's LVCSR system. Sonic is a continuous-density HMM (CDHMM) based recognizer. The acoustic models are decision-tree state-clustered HMMs that use associated Gamma probability density functions to model state-durations. The 39 dimensional feature vector contains 12 statics, deltas and delta-deltas along with energy, delta and delta-delta energy. We used pre-emphasis and a Hamming window of length $25ms$ and a skip rate of $10ms$ on the frame data before further processing. Cepstral mean normalization (CMN) was also utilized on the final feature vectors. 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 task is an extended digit recognition task using the CU-Move database [7, 21, 8], in which a total of 60 speakers balanced across gender and age (18-70 yrs.) were in the training set. The test set contained 77 speakers. The model set had 450 models and $10K$ Gaussians [18]. The vocabulary size was 42 including silence (SIL) and unknown word (UNK). The dictionary is very convenient for telephone dialing applications since it contains most necessary words like "dash", "pound", "sign" in addition to numbers. The results are given in Table 1. The relative improvement of PMVDR over MFCC is 35.5% for female speakers and 19.0% for male speaker, an average of 27.3%. These values are 23.8%, 14.2% and 18.8% for PMCC. The improvements can be attributed to the elegant spectral properties of MVDR and accurate perceptual warping.

Table 1: *WERs[%] for CU-Move task with different front-ends.*

Gender/Sys.	MFCC	PMCC	PMVDR
$Female$	10.41	8.81	6.71
$Male$	12.12	11.45	9.82
Overall	**11.19**	**10.02**	**8.14**

The improvement is remarkable especially for *female* speakers. This supports the claim that MVDR methodology is especially effective for high-pitch speech. The parameters, i.e. the warping factor, α, and LP order, Q, were tuned on a 17-speaker development set. A good set of values are $\alpha = 0.42$ and $Q = 22$. An interesting observation is the relatively high value of LP order. The typical value of LP order for LP-based cepstral coefficients (LPCCs) is $Q = 13$ while we need $Q = 22$ for PMVDR. This can be attributed to the smoother nature of MVDR spectrum, i.e. we need a higher order to model just enough detail necessary for accurate recognition. We note that, a more exhaustive optimization should be performed on a large vocabulary task.

5. Noise and Speaker Robustness

For a noisy database, such as CU-Move, identifying the sources of improvement is rather difficult. However, there are two possible areas we consider, namely *robustness to additive noise* and *robustness to speaker variability*. Channel effects are ignored because, in our case, the recognition is done within the vehicle so the audio data need not be sent over a channel. We used the same CU-Move test set in both analyses.

5.1. Robustness to Additive Noise

Obtaining acceptable recognition performance in noise is a desirable property of a feature extraction scheme. We believe that an analysis should be performed, rather then citing the final recognition results, to prove this claim since there might be other merits of the new feature extraction scheme that could affect the performance in a positive manner.

In order to perform the noise robustness analysis, we propose to use *Segmental SNR (SSNR)* [22] versus *word error rate* (WER) [7]. 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 [18] to align the data and extract speech-silence segmentation from the phone alignments.

2. Use NIST's SSNR utility [22] to compute SSNR for each utterance. The SSNR calculation utility produces an accurate enough SNR estimate for our purpose.

3. Average the SSNR for each speaker and generate a scatter plot of the SSNR vs. WER for the entire test set.

The resulting plot is a measure of dependency between the SSNR and WER. We propose to use the *correlation coefficient*, q, to evaluate the degree of dependency. For a truly noise robust feature extraction scheme, the correlation of SSNR and WER should be close to 0. *The smaller the correlation coefficient, the less the degree of dependency.*

We performed the analysis outlined above for three different acoustic modeling strategies analyzed in this paper, namely MFCC, PMCC and PMVDR. The resulting scatter plot is given in Figure 3. Note the highly varying nature of SSNR (as much as 10 dB) across the speakers. This variation is a typical property of real car environments since there is a wide range of both SNRs and spectral structures for the noise. The correlation coefficients are summarized in Table 2. There is a negative correlation between the SSNR and WER, as it should be since as the SSNR increases (data becomes less noisy) we would expect the WER to drop. 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 least robust modeling strategy for noise robust in-vehicle speech recognition is MFCC, while PMVDR delivers the greatest level of robustness to noise amongst the three. Thus, we have shown that PMVDR feature extraction methodology is indeed less susceptible to additive noise than MFCC and PMCC.

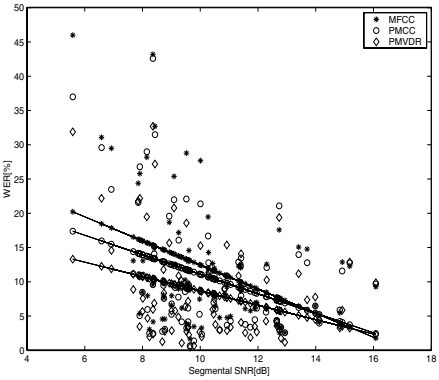

Figure 3: *Scatter plots and* 1^{st} *order fits for MFCC, PMCC, and PMVDR*

Table 2: *Corr. coefs. of SSNR and WER for the 3 front-ends.*

Measure/Systems	MFCC	PMCC	PMVDR
Correlation Coef.	-0.407	-0.355	-0.309

5.2. Robustness to Speaker Variability

We also claim that the PMVDR scheme better suppresses speaker dependent information than MFCC and PMCC. This section aims to evaluate the three feature extraction schemes in terms of their robustness to speaker variability. We use a modified Linear Discriminant Analysis (LDA) scheme proposed in [23] to evaluate the robustness to speaker variability. The scheme is basically a modified LDA in which we compute the within-class scatter matrix with respect to speaker variability; therefore, the LDA objective function is now optimized with respect to speaker variations on phone classes. For computational details, we refer the reader to [23, 24].

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 [23]. A good measure of this property is the *trace* of $S_W^{-1}S_B$. Here, S_W refers to the within-class scatter matrix while S_B denotes the between-class scatter matrix. The trace is the sum of the eigenvalues λ_i of $S_W^{-1}S_B$[23]. Now we can define a measure for assessing speaker variability, the *trace measure*. The interpretation for the trace measure is that the trace equals the sum of the variances in the principal directions. It can also be 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 the higher the class separability, the lower the recognition error rate [23]. The trace measure is formulated below and used as a measure of inter-speaker variability within phonemes in this study.

$$Tr(d) = \sum_{i=1}^{c} \lambda_i \tag{5}$$

Where c is the number of classes for which we have data. In Table 3, we give the trace measure for MFCC, PMCC and PMVDR feature extraction schemes. We conclude from the table that PMVDR shows less speaker variability.

Table 3: *Trace measure for different front-ends.*

Measure/Systems	MFCC	PMCC	PMVDR
Trace	33.17	34.38	34.58

6. Conclusions

In this paper, we proposed a novel algorithm to compute cepstral coefficients to represent speech for robust in-vehicle speech recognition. We incorporate perceptual considerations directly on the FFT power spectrum and utilize the MVDR spectrum as the spectral envelope estimation technique. The envelope is encoded into the cepstral coefficients to have an uncorrelated representation. The resulting coefficients, PMVDRs, are shown to outperform conventional MFCCs and recently proposed PMCCs on an extended digit recognition task in the car. The PMVDR is proven by two separate analyses to be less susceptible to additive noise and to be more efficient in suppressing speaker dependent information that exists in the spectrum. Thus, the PMVDR is an effective candidate to replace MFCC in future state-of-the-art speech recognition systems working in noisy environments.

7. Acknowledgments

The authors would like to thank Dr. S. Dharanipragada of Human Language Technologies at IBM for his very fruitful discussions.

8. References

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

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

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

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

[5] Hermansky, H., "Perceptual Linear Prediction (PLP) Analysis of Speech" *J. Acoust. Soc. Am., pp 1738-1752, 1990.*

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

[7] Yapanel, U., Zhang, X., Hansen, J. H. L., "High Performance Digit Recognition in Real Car Environments", *Proc. ICSLP'02*

[8] http://cumove.colorado.edu

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

[10] Murthi, M. N. and Rao, B. D., "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech & Audio Proc., May 2000.*

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

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

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

[14] Tokuda, K., Masuko, T., Kobayashi, T., Imai, S., " Mel-generalized Cepstral Analysis-A Unified Approach to Speech Spectral Estimation", *Proc. ICSLP'94*

[15] Smith, J. O. and Abel, J. S., " Bark and ERB Bilinear Transforms", *IEEE Trans. Speech & Audio Proc., Nov. 1999*

[16] Makhoul, J., "Linear Prediction: a Tutorial Review", *Proc. of IEEE, pp.561-580, 1975*

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

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

[19] Pellom, B. and Hacioglu, K., "Recent Improvements in the CU Sonic ASR System for Noisy Speech: The SPINE Task," *ICASSP'03*

[20] http://cslr.colorado.edu

[21] Hansen, J. H. L., Angkititrakul P., Yapanel, U. et. al., "CU-Move: Analysis & Corpus Development for Interactive In-vehicle Speech Systems", *Proc. Eurospeech'01*

[22] SPHERE software package, www.nist.gov

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

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