MVDR BASED FEATURE EXTRACTION FOR ROBUST SPEECH RECOGNITION

Satya Dharanipragada

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

Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego
San Diego, CA 92093-0407.
brao@ece.ucsd.edu

ABSTRACT

This paper describes a robust feature extraction method for continuous speech recognition. Central to the method is the Minimum Variance Distortionless Response (MVDR) method of spectrum estimation and a feature trajectory smoothing technique for reducing the variance in the feature vectors. The above method, when evaluated on continuous speech recognition tasks in a stationary and moving car, gave an average relative improvement in WER of greater than 30%.

1. INTRODUCTION

Estimating the time-varying spectrum is a key first step in most feature extraction methods for speech recognition. Cepstral coefficients derived from a modified short-time spectrum is the most popular feature set and has been empirically observed to be the most effective for speech recognition. The modification of the spectrum is often based on perceptual considerations. Mel-Filtered Cepstral Coefficients (MFCC) is one such popular feature set.

Both parametric and nonparametric methods of spectrum estimation have been studied for speech modeling. Of the parametric methods the LPC based all-pole spectrum is most widely used. However, it has been noted, in the speech modeling literature, that for medium pitch voiced speech and high pitch voiced speech, LP based all-pole models do not provide good models of the spectral envelope, [1]. Furthermore, LP based cepstra are known to be very sensitive to noise. Nonparametic spectrum estimation methods such as the FFT-based Periodogram or Modified Periodogram on the other hand are attractive since these methods are entirely data-independent and hence do not suffer from problems arising due to modeling deficiencies. However, these methods often are not robust and therefore perform poorly in noisy and adverse conditions. In general, parametric methods with accurate models suited for the given application should be able to provide more accurate and robust estimates of the short-term power spectrum.

In this paper, we examine the use of the recently proposed Minimum Variance Distortionless Response (MVDR) spectrum-based modeling of speech, [2], for speech recognition. In [2], it was shown that high order MVDR models provide elegant envelope representations of the short-term spectrum of voiced speech. This is particularly suited for speech recognition where model order is not a concern. Furthermore, it was shown that 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.

In speech recognition, in addition to faithful representation of the spectral envelope, statistical properties such as the bias and variance of the spectral estimate are of great interest too. Variance in the feature vectors has a direct bearing to the variance of the Gaussians modeling the speech classes. In general, reduction in feature vector variance increases class separability. Improved class separability can potentially increase recognition accuracy and decrease search speed. We present a simple smoothing technique that effectively reduces variance of the feature vectors and therefore the Gaussians modeling of the speech classes.

2. MVDR BASED FRONTEND

In nonparametric spectrum estimation methods like the FFT-based Periodogram method, the power is measured using a single sample at the output of a bandpass filter centered at the frequency of interest [3, 4]. The nature of the bandpass filter is frequency and data independent, and determined only by the nature and length of the window used. The window length is usually equal to the data segment length. For speech recognition we are more interested in the statistical stability of the estimate than the spectral resolution limit. Two statistical properties of the spectrum estimate are of interest, viz., the bias and variance. A large bias or variance in estimates will ultimately lead poor acoustic models. Bias is mainly caused by the leakage of power from surrounding frequencies through the sidelobes or the mainlobe of the bandpass filter. Since a single sample is used to estimate the power, Periodogram estimates have a large variance. Furthermore, since the bandpass filter is data indepen-

dent there is no flexibility to modify the sidelobe properties to suppress dominant neighboring peaks. An approach to lower the variance is to use the Modified Periodogram or the Welch method. Such an approach leads to lower variance at the expense of larger bias. The larger bias is a consequence of the smaller window length resulting in a bandpass filter with larger bandwidth. Also the bandpass filter is data independent. Both these shortcomings will be addressed by the MVDR and variance reduction methods described next.

2.1. Bias and Variance Reduction

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 [6, 5]

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

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 (c.f. Section 2.2). Futhermore, smaller filter lengths, for the same bias and variance, enables usage of a second temporal averaging technique for further variance reduction in the feature vectors, as will be explained in the Section 2.4.

2.2. MVDR Spectrum Computation

Fortunately, as in the FFT based methods, in the MVDR method there is no explicit need to design a separate filter $h_l(n)$ for each frequency ω_l. In fact, the MVDR spectrum for all frequencies can be conveniently represented in a parametric form. It can be shown that the output power of the optimum constrained filter, and hence the MVDR spectrum for all frequencies can be simply computed as [6]

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

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 [4]. For computational purpose, 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 [6, 5]

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

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. The minimum-phase MVDR all-pole filter $1/B(z)$, if needed, can be obtained by a spectral factorization. The MVDR all-pole filter $1/B(z)$ is stable and causal, and can be used in a manner similar to the way in which LP filters are used in speech processing systems.

2.3. Mel-Cepstra Computation

There are two possible approaches to computing the cepstrum from the MVDR spectrum. One alternative is to compute the all-pole model and derive the cepstra directly from the coefficients of the all-pole filter $B(z)$. The other alternative is to compute the spectrum from the MVDR polynomial using the FFT and then compute the cepstral coefficients from the spectrum in the standard way. In this paper, we choose the second alternative because of the ease with which perceptual considerations can be incorporated.

2.4. A Second Variance Reduction Step

The basic idea behind the second variance reduction step is smoothing. To understand this, consider the following example. Let $x_1, x_2, \cdots, x_P$ be P uncorrelated random variables with zero mean and variance σ^2. Consider, $y = \frac{1}{P}\sum_{i=1}^{P} x_i$. Clearly, y has zero mean and variance $\frac{\sigma^2}{P}$. Thus an estimate obtained by averaging P uncorrelated estimates provides a factor of P reduction in variance.

In the context of the speech recognition frontend, smoothing can be performed either to the power spectral samples

or to the MFCC. We chose to smooth the MFCC in our experiments. Averaging the MFCC is equivalent to taking a geometric mean of the spectral samples. In order to obtain several uncorrelated estimates of the MFCC one needs data segments that are uncorrelated with each other. For a WSS process with a sharply decaying correlation function, data segments that are sufficiently separated temporally will be uncorrelated. Thus, by splitting the data segment into several overlapping segments and computing power spectral estimates from each segments we can obtain power spectral estimates that are reasonably uncorrelated. The MVDR estimation method facilitates this further because it requires shorter filter lengths for the same bias and variance. This effectively lets us create more uncorrelated data segments from a given frame of speech samples.

Therefore, instead of generating a single MFCC vector from a frame of speech, samples from the start of the current frame to the start of the next frame are split into several overlapping segments and an MFCC vector is computed from each segment. These vectors are then averaged to get the smoothed MFCC vector for that frame. This is equivalent to generating feature vectors at a high frame-rate and downsampling the resulting trajectories after low pass filtering in the time domain. The filtering operation is performed by simple averaging. This approach of filtering, motivated purely from statistical stability considerations, is very different from RASTA processing, [7], which is motivated from human auditory perception considerations. Furthermore, the filtering, here, is done within each frame and not across frames like in RASTA. Figure 1 shows a schematic diagram of the MVDR based front-end processor.

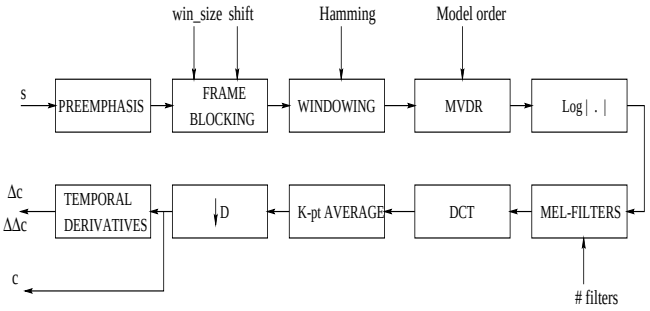

Figure 1: Schematic diagram of the MVDR-based front-end processor

3. EXPERIMENTAL RESULTS

We experimented with this new feature extraction technique in a speech recognition system for a voice-activated car navigation system. The training data consists of a combination of cellular, speaker-phone and car data collected using an appropriately placed microphone in a car. Car noise at various speeds was collected using a microphone over a cellular channel. Both clean speech and noise-added speech was used to train the systems.

3.1. System Description

All experiments were conducted on the IBM rank-based LVCSR system. The IBM LVCSR 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 [8]. The training 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.

3.2. Experimental Set-up

The baseline system was trained using standard FFT-based MFCC vectors. Speech was coded into 25 ms frames, with a frame-shift of 10 ms. Each frame was represented by a 39 component vector consisting of 13 MFCCs and their first and second time derivatives. Overall, the decision tree had 2615 leaves. Each leaf had 15 Gaussian mixture components for the output distribution.

Next, 13 dimensional MFCC features were generated at a high rate of 500 frames/s (frame-shift of 2ms) using the MVDR spectrum estimate. A model order of 60 was chosen for the LPC analysis. Since we are dealing with car noise, the 24 traingular Mel-filters were chosen in the frequency range [200Hz — 3800Hz]. A smoothed MFCC stream was generated by taking a 5-point average and downsampling by a factor of 5 to produce a 100 frames/sec stream. First and second time derivatives are then computed from the smoothed MFCC stream. With this new feature stream, the means and the variances of the Gaussians and the transition probabilities of the HMM's were re-estimated using a Baum-Welch procedure.

3.3. Results

Figure 2 shows a histogram of the ratios of the variance of the Gaussians in the baseline (FFT-based) system and the variance of the Gaussians after retraining with the MVDR-based MFCC. The large mass at 0.5 clearly indicates a strong reduction in the variances of the re-estimated Gaussians.

For the test set, several speakers were recorded in a stationary and moving car at 30 mph and 60 mph. Ten different sub-tasks within a navigation task, each with a different vocabulary size, were used to create a test set. Simple BNF grammars were constructed for each task and were used to guide the search. Tables 1, 2, and 3 give a detailed comparison of the word error rates with the FFT-based MFCC system and the new MVDR-based MFCC system. Results

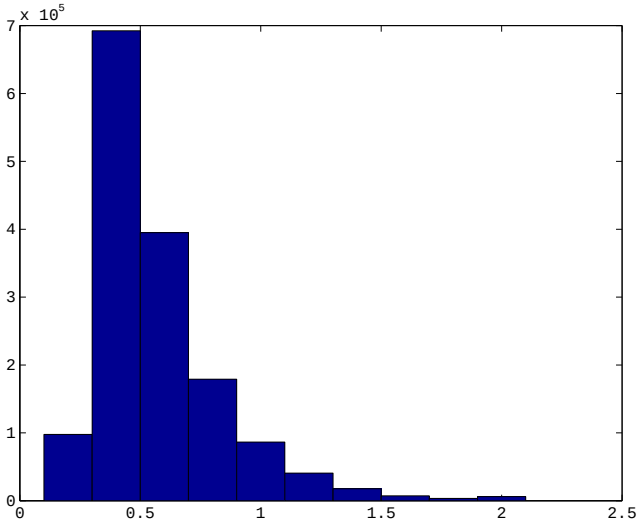

Figure 2: Histogram of $\frac{\sigma^2_{mvdr}}{\sigma^2_{base}}$

Task	VocSize	#words	Baseline	MVDR
airports	335	750	12.13	7.33
banks	63	985	9.64	5.69
commands	22	439	11.85	13.44
county	1876	194	45.36	29.38
gas-stations	16	101	2.97	1.98
hotels	55	461	6.29	3.90
reactions	33	189	12.70	5.29
service stations	39	164	7.32	2.44
US cities	12000	227	52.86	45.81

Table 1: WER of FFT-based baseline versus MVDR-based MFCC system at 0 mph

clearly indicate a significant improvement in the recognition accuracy in all the tasks and under all conditions. Average relative improvements of 27.9%, 32.3%, 38.5% were observed in the 0 mph, 30 mph, and the 60 mph conditions respectively.

4. CONCLUSIONS

We described a robust feature extraction method for continuous speech recognition. The method uses the MVDR spectrum estimation technique and a variance reduction technique based on the temporal smoothing of the cepstral trajectories. The above method gave very significant improvements in word error rate on continuous speech recognition tasks in a stationary and moving car environments.

5. REFERENCES

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

Task	VocSize	#words	Baseline	MVDR
airports	335	672	8.48	6.10
banks	63	898	4.90	2.12
commands	22	346	14.74	9.83
county	1876	189	49.21	43.92
gas-stations	16	83	1.20	1.20
hotels	55	378	2.12	1.85
reactions	33	157	15.29	5.73
service stations	39	142	8.45	0.00
US cities	12000	228	50.88	35.53

Table 2: WER of FFT-based baseline versus MVDR-based MFCC system at 30 mph

Task	VocSize	#words	Baseline	MVDR
airports	335	378	16.93	7.14
banks	63	475	8.84	4.21
commands	22	178	15.17	11.80
county	1876	86	62.79	46.51
gas-stations	16	47	14.89	2.13
hotels	55	218	9.63	7.34
reactions	33	87	17.24	6.90
service stations	39	76	7.89	7.89
US cities	12000	111	70.27	50.45

Table 3: WER of FFT-based baseline versus MVDR-based MFCC system at 60 mph

[2] 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.

[3] P.D. Welch, "The use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short Modified Periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 70–76, June 1967.

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

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

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

[7] H. Hermansky and N Morgan, "RASTA Processing of Speech," *IEEE Trans. on Speech and Audio Processing,* Vol. 2, No. 4, pp. 587–589, October 1994.

[8] L.R. Bahl, P.V. deSouza, P.S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny, "Robust methods for context-dependent features and models in a continuous speech recognizer," *Proc., Intl Conf. on Acoust., Speech, and Sig. Proc.*, 1994, pp. I-533–536.