SPECTRAL ESTIMATION OF VOICED SPEECH USING A FAMILY OF MVDR ESTIMATES

Rajesh M. Hegde, Yuzhe Jin, and Bhaskar D. Rao

Department of Electrical and Computer Engineering,
University of California San Diego, La Jolla, CA 92039-0407, USA
{rhegde,yujin,brao}@ucsd.edu

ABSTRACT

We present a robust approach to modeling voiced speech using a family of Minimum Variance Distortionless Response (MVDR) spectral estimates. The method exploits the fact that for a fixed model order, for a sinusoidal signal in noise, the MVDR estimate at the sinusoidal frequency is approximately related to the sinusoidal and noise power in a simple linear manner with the coefficients being dependent on the model order. Modeling voiced speech as a sum of harmonic signals, we then use the aforementioned relationship along with a least squares approach to combine a family of MVDR estimates (MVDR estimates of different orders) and develop a robust approach for modeling voiced speech. Experimental results of spectral estimation of sinusoids, synthetic vowels, and actual speech signals at SNR of 0 dB and 5 dB using this approach indicate an increased resolution in the estimated MVDR spectra. The MFCC computed from the MVDR spectra using this approach are also used for speaker identification experiments on the TIMIT database at various SNR. The results indicate a reasonable improvement in recognition performance when compared to the MFCC and the fixed order MVDR-MFCC.

Index Terms— Spectrum estimation, Minimum variance distortionless response spectrum (MVDR), MFCC, Spectral distortion.

1. INTRODUCTION

In this paper, we propose a novel technique to estimate the spectrum of voiced speech signals in additive noise using the minimum variance method of spectrum estimation as a building block. The minimum variance distortionless response (MVDR) [1, 2], method of spectral estimation also called the Capon method has been used to represent short time envelope of speech in [3, 4]. It is also effective in modeling unvoiced and mixed speech spectra reasonably well. In [5], the estimation of the Capons maximum likelihood spectra of pure sinusoids and sinusoids in noise has been discussed. In particular, interesting insights are provided on the family of spectral estimates, i.e. behavior as the model order is increased. In the past work, often a suitable compromise model order is selected and the spectral estimate is computed. In contrast, in this work we use a family of estimates (estimates obtained using different model orders), along with the structure of voiced speech in a synergistic manner to provide a robust estimate of the harmonic content. We exploit the fact that for a fixed model order, for a sinusoidal signal in noise, the MVDR estimate at the sinusoidal frequency is approximately related to the sinusoidal and noise power in a simple linear manner with the coefficients being dependent on the model order. Modeling voiced speech as a sum of harmonic signals, we then use the aforementioned relationship and a least squares approach to combine a family of MVDR estimates (MVDR estimates of different orders) and develop a robust approach for modeling voiced speech. As a proof of concept, the results of spectral estimation of sinusoids, synthetic vowels, and actual speech signals in noise are illustrated. We also compare spectral distortions of the MVDR spectrum estimated from the proposed robust approach, to both the conventional MVDR and the DFT based techniques in terms of the average error distributions. The results indicate that the average deviation of the noisy speech signal spectrum from the clean speech signal spectrum is the least for this approach, when compared to conventional MVDR and DFT based methods. MFCC are extracted from the MVDR spectrum computed using the proposed approach and are called the CF-MVDR-MFCC. These features are used as the front end to build a speaker identification system using the TIMIT database [6]. Results of performance evaluation indicate a reasonable improvement at SNR of 0dB and 5dB.

This work was funded by the National Science Foundation under award numbers 0331707, and 0331690, and UC MICRO grant numbers 05-033, and 06-174.

2. THE MINIMUM VARIANCE SPECTRUM ESTIMATION

The MVDR spectrum estimate [1, 2], is a non parametric, data adaptive technique that can be used to obtain better resolution than the DFT based spectrum estimation methods. The MVDR spectral estimate of order M is given by

$$R_M^{mvdr}(e^{j\omega}) = \frac{1}{\mathbf{v}^H(\omega)\mathbf{R}_x{}^{-1}\mathbf{v}(\omega)}, \tag{1}$$

where $\mathbf{R}_x$ is the $(M) \times (M)$ data autocorrelation matrix and

$$\mathbf{v}(\omega) = [1, e^{j\omega}, e^{j2\omega}, e^{j3\omega}, ..., e^{j(M-1)\omega}]^T. \tag{2}$$

This estimate has some interesting properties which we briefly mention below. It can be efficiently computed exploiting the relationship with linear prediction methods as

$$R_M^{mvdr}(e^{j\omega}) = \frac{1}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}} \tag{3}$$

where the parameters $\mu(k)$, are obtained by a simple non iterative computation involving the linear prediction coefficients [1]. The correlation matrix is estimated from the data and in our work it is computed using the forward backward procedure. Conceptually, the filter bank interpretation is most insightful for our problem. The MVDR spectrum at a given frequency ω_k can be viewed as the power at the output of a FIR filter described by its coefficients $\beta = [h(0), h(1), ..., h(M-1)]^T$, which are obtained by solving the following constrained optimization problem

$$\min_{\beta} \beta^H \mathbf{R}_x \beta \;\text{ subject to } \beta^H \mathbf{v}(\omega) = 1.$$

The linear constraint ensures the signal of interest is not distorted and the minimization of the output power minimizes leakage from other frequencies.

3. THE MINIMUM VARIANCE SPECTRUM ESTIMATION FOR AN EXPONENTIAL SIGNAL IN NOISE

The method developed exploits an important property of the MVDR spectral estimate for data consisting of sinusoidal signals in noise. To develop the insight, we first consider a data sequence $x[n]$ which consists of a single undamped complex exponential signal with frequency ω_k, corrupted with additive white noise $w(n)$, i.e.

$$x(n) = \Psi_k e^{j\omega_k n} + w(n) \tag{4}$$

where

$$\Psi_k = |\Psi_k| e^{j\psi_k} \tag{5}$$

and $|\Psi_k|$ is the spectral magnitude and ψ_k is the uniformly distributed random phase. The correlation matrix of the signal x(n) is given by

$$R_x = |\Psi_k|^2 \mathbf{v}^H(\omega_k)\mathbf{v}(\omega_k) + \sigma_w^2 \mathbf{I} \tag{6}$$

Using the matrix inversion lemma to compute the inverse of the matrix R_x, and then substituting in Equation 1, we have the minimum variance estimate of an exponential signal in noise at frequency ω_k as

$$R_M^{mvdr}(e^{j\omega}) = \frac{\sigma_w^2}{M - \frac{|\Psi_k|^2}{\sigma_w^2 + M|\Psi_k|^2}|\mathbf{v}^H(\omega)\mathbf{v}(\omega_k)|^2} \tag{7}$$

From Equation 7, since the norm of the frequency vector $\mathbf{v}(\omega)$ is equal to M, the minimum variance spectral estimate of an exponential in noise at the frequency ω_k, reduces to

$$R_M^{mvdr}(e^{j\omega_k}) = \frac{\sigma_w^2}{M} + |\Psi_k|^2 \tag{8}$$

Hence for a frequency ω_l, $\omega_l \neq \omega_k$, and reasonably far from ω_k, when $M >> 1$, we have

$$\mathbf{v}^H(\omega_k)\mathbf{v}(\omega_l) \approx 0 \tag{9}$$

From Equations 8 and 9, we have an MVDR spectral estimate of noise as

$$R_M^{mvdr}(e^{j\omega_l}) = \frac{\sigma_w^2}{M} \tag{10}$$

Note that the above analysis assumed a single complex exponential signal. If there are more than one undamped complex exponentials, then if we have a sufficiently large model order, Equation 8 can be expected to be approximately true at the frequencies corresponding to the exponentials. This can be understood from the filter bank interpretation. The output power minimization will attempt to minimize the contributions from the other exponentials making the relationship to still hold in an approximate manner [4].

4. MODELING USING A FAMILY OF MVDR SPECTRAL ESTIMATES

Sinusoidal modeling has often been used to model voiced speech [7]. We model a frame of voiced speech as a sum N complex exponentials with the frequencies ω_k being a integer multiple of the fundamental (pitch) frequency ω_0. Such a signal is described by

$$x(n) = \sum_{k=0}^{N-1} \Psi_k e^{j(\omega_k n + \psi_k)} \tag{11}$$

where $k = \{0, 1, 2, ...(N-1)\}$, are the frequency components of the signal. Assuming that the frequency components are known, the MVDR estimate can be computed at each of these frequencies using varying model orders, $M = M_1, M_2, ..., M_L$. For a given harmonic frequency ω_k, using Equation 8, we have

$$\begin{aligned} R_{M_1}^{mvdr}(e^{j\omega_k}) &= \tfrac{\sigma_w^2}{M_1} + |\Psi_k|^2 + \epsilon_1 \\ R_{M_2}^{mvdr}(e^{j\omega_k}) &= \tfrac{\sigma_w^2}{M_2} + |\Psi_k|^2 + \epsilon_2 \\ ..... &= ..... + ... + ... \\ R_{M_L}^{mvdr}(e^{j\omega_k}) &= \tfrac{\sigma_w^2}{M_L} + |\Psi_k|^2 + \epsilon_L \end{aligned} \tag{12}$$

where $\epsilon_1, \epsilon_2, ..., \epsilon_L$, are introduced to model the various sources of error, e.g correlation matrix estimation error, leakage from other components etc. The above set of equations for a single frequency ω_k can be written in matrix form as

$$Y = AX + \epsilon \tag{13}$$

where $Y = [R_{M_1}^{mvdr}(e^{j\omega_k}), R_{M_2}^{mvdr}(e^{j\omega_k}), ..., R_{M_L}^{mvdr}(e^{j\omega_k})]^T$, $X = [\sigma_w^2, |\Psi_k|^2]^T$, and A is given by

$$A = \begin{pmatrix} 1 & M_1 \\ 1 & M_2 \\ . & . \\ . & . \\ . & . \\ 1 & M_L \end{pmatrix}$$

Equation 13, naturally suggests a linear least squares solution for X. The matrix X, can be estimated by minimizing the weighted norm of the error, $||e||_W^2 = ||Y - AX||_W^2$. The weighted least square solution to X is given by

$$X = (A^H W A)^{-1} A^H W Y. \tag{14}$$

The matrix W is a diagonal matrix of weight vectors which can be computed based on the knowledge that the error in the MVDR spectral estimation process. For example, lower model orders can be assumed to have more error than higher model orders and this can be reflected in the choice of W. For simplicity, in this work we assume $W = I$. In the discussions that follow, we only illustrate the spectral estimation of $|\Psi_k|^2$, which essentially is the focus of this work. Since the proposed estimate is obtained by **C**ombining a **F**amily of MVDR spectral estimates, we refer to this estimate as the CF-MVDR estimate in the ensuing discussions.

4.1. CF-MVDR spectrum estimation of sinusoids in noise

To demonstrate the effectiveness of the method developed, we start our experimental analysis of the CF-MVDR spectrum estimation by considering a signal composed as a set of two sinusoids.

$$u(n) = \sum_{k=1}^{2} c_k cos(\omega_k n + \psi_k) \tag{15}$$

where ω_k = 0.3 and 2, is the frequency in radians, and c_k = 1, is the amplitude of the sinusoids. White noise is scaled and added to the signal u(n) to realize SNRs of 0 dB and 5 dB. In Figures 1 (a) and (b) respectively, are shown, the 15^{th} order MVDR spectral estimate and the CF-MVDR spectral estimate, with a range of model order 10-15, for a SNR of 0 dB. While in Figures 1 (c) and (d), similar plots for a SNR of 5 dB are illustrated. The CF-MVDR estimate shows some interesting properties like bandwidth normalization, theoretically similar to a very high order MVDR spectrum.

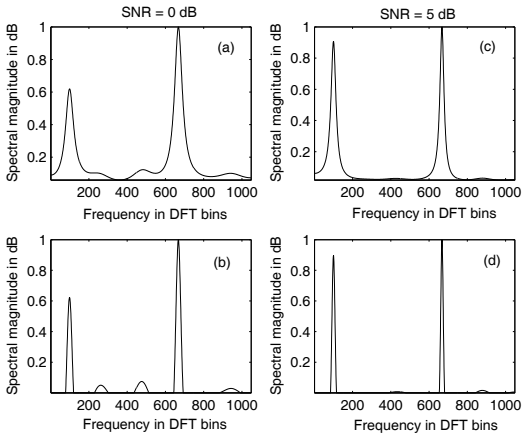

Fig. 1. Illustration of MVDR and CF-MVDR spectrum estimation of sinusoids in noise. (a) MVDR spectrum of 2 sinusoids at SNR of 0 dB, (b) CF-MVDR spectrum of 2 sinusoids at SNR of 0 dB, (c) MVDR spectrum of 2 sinusoids at SNR of 5 dB, and (d) CF-MVDR spectrum of 2 sinusoids at SNR of 5 dB.

4.2. CF-MVDR spectrum estimation of synthetic vowels in noise

To generate a synthetic vowel, we use the system function

$$H(z) = \frac{1}{1 - 2e^{-\pi B_i T} cos\omega_i T z^{-1} + e^{-2\pi B_i T} z^{-2}} \qquad (16)$$

where ω_i corresponds to the formant, B_i to the bandwidth and T to the sampling period. Using Equation 16, we generate a synthetic vowel with the following values F1 = 500Hz, F2 = 1200Hz, B_i = 10% of F_i, and T = 0.000125 s corresponding to a sampling rate of 8 KHz. The synthetic vowel is generated using a pulse train where each pulse is separated apart by $1/60$ sec. and therefore has a pitch frequency of 60 Hz. White noise is added as described in the previous Section. Figures 2 (a) and (b) illustrate the the 15^{th} MVDR spectral estimate and the CF-MVDR spectral estimate, with a range of model order 10-15, of the synthetic vowel for a SNR of 0 dB respectively. In Figures 2 (c) and (d), similar plots for a SNR of 5 dB are illustrated. The CF-MVDR estimate as shown in the figures indicate some interesting properties like high resolution and low bandwidth spectrum estimation for vowels in noise. It is interesting to note that high resolution of the CF-MVDR also resolves the second formant of the vowel into two, which can be eliminated by fixing the correct range of model order as indicated by our experiments.

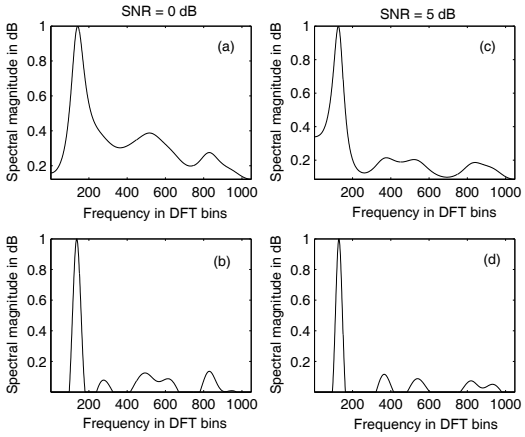

Fig. 2. Illustration of the MVDR and CF-MVDR spectrum estimation of a synthetic vowel in noise. (a) MVDR spectrum of synthetic vowel at SNR of 0 dB, (b) CF-MVDR spectrum of synthetic vowel at SNR of 0 dB, (c) MVDR spectrum of synthetic vowel at SNR of 5 dB, and (d) CF-MVDR spectrum of synthetic vowel at SNR of 5 dB.

4.3. CF-MVDR spectrum estimation of speech

We illustrate the CF-MVDR spectrum estimation of actual speech signals, by considering a word *matlab* uttered by a female speaker sampled at 8 KHz. In Figures 3 (a) and (b), the 80^{th} order MVDR spectral estimate and the CF-MVDR spectral estimate, with a range of model order 50-80, for a SNR of 0 dB are shown as spectrogram plots. While in Figures 3 (c) and (d), similar plots for a SNR of 5 dB are illustrated. The CF-MVDR spectrograms illustrated in Figure 3, also displays some interesting properties like high resolution and optimal bandwidth spectrum estimation for actual speech signals at low SNR.

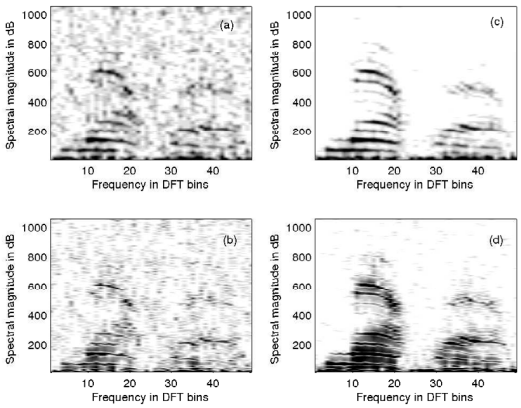

Fig. 3. Illustration of the MVDR and CF-MVDR spectrum estimation of actual speech. (a) MVDR spectrogram of speech at SNR of 0 dB, (b) CF-MVDR spectrogram of speech at SNR of 0 dB, (c) MVDR spectrogram of speech at SNR of 5 dB, and (d) CF-MVDR spectrogram of speech at SNR of 5 dB.

5. PERFORMANCE EVALUATION

In this Section we evaluate the CF-MVDR spectral estimation technique in terms of its robustness and also list some recognition results for a speaker identification task at different SNRs.

5.1. Robustness of CF-MVDR spectrum estimation

We compare the robustness of the MVDR, CF-MVDR and the conventional DFT based spectrum estimation techniques in the presence of white noise at different values of SNR using average error distributions (AED). Thirty utterances from different dialect regions, consisting of both female and male speakers, from the TIMIT database

[6] are picked for the analysis. White noise scaled by a factor is added and the corresponding SNR computed. The average error distributions between the clean and the noisy speech across all frames corresponding to the 30 sentences are then calculated for two different values of SNR 0 and 5 dB. Figures 4 (a), (c), and (e) correspond

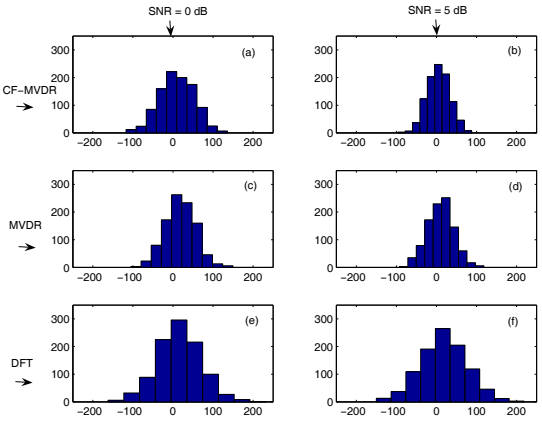

Fig. 4. (a) AED for CF-MVDR spectral estimation at SNR of 0 dB, (b) AED for CF-MVDR spectral estimation at SNR of 5 dB,, (c) AED for MVDR spectral estimation at SNR of 0 dB, (d) AED for MVDR spectral estimation at SNR of 5 dB, (e) AED for DFT based spectral estimation at SNR of 0 dB, and (f) AED for DFT based spectral estimation at SNR of 5 dB.

to the average error distribution of the CF-MVDR, MVDR, and DFT based spectrum estimation respectively for a SNR of 0 dB, while Figures 4 (b), (d), and (f) are similar plots at SNR of 0, and 5 dB respectively. It is clear from Figure 4, that average deviation of the noisy speech cepstra from the clean speech cepstra is the least for the CF-MVDR technique when compared to the MVDR and DFT based spectral estimation techniques.

5.2. Experimental results for speaker identification

The baseline system used in this study uses the principle of likelihood maximization. A series of GMMs are used to model the voices of speakers for whom training data is available. Single state, 64 mixture Gaussian mixture models (GMMs) are trained for each of the 100 speakers picked up across the various dialect regions from the the TIMIT database [6]. A classifier evaluates the likelihoods of the unknown speaker's voice data against these models. The model that gives the maximum accumulated likelihood is declared as the correct match. Out of the 10 sentences for each speaker, 8 were used for training, and 2 were used for testing. The tests were conducted on 100 speakers (100 x 2 tests) and the number of tests was 200. Results of performance evaluation for various features on the TIMIT [6] at various SNR are listed in Table 1. MFCC are computed from the CF-MVDR (model order range of 10-15), MVDR (15^{th} order), and the DFT power spectrum and are called CF-MVDR-MFCC, MVDR-MFCC and the MFCC respectively. The CF-MVDR-MFCC indicate a reasonable improvement in recognition performance over Fixed order MVDR-MFCC and the MFCC.

Table 1. Recognition performance of various features for speaker identification.

Front end	Signal to Noise Ratio	% Error Rate
CF-MVDR-MFCC	Clean	0.5%
	10 dB	2%
	5 dB	16%
	0 dB	42%
Fixed Order MVDR-MFCC	Clean	0.5%
	10 dB	3%
	5 dB	18%
	0 dB	45%
MFCC	Clean	0.5%
	10 dB	5%
	5 dB	21%
	0 dB	46%

6. CONCLUSION

The significance and applications of the fixed order MVDR spectrum estimation in speech processing has been discussed in earlier efforts. This work presents a robust method for modeling voiced speech by using a least squares approach to combine a family of MVDR spectrum estimates. This approach also indicates reasonable improvements both in the estimation of speech spectra and in speaker identification experiments conducted on the TIMIT database. The computation of the weighting factors based on the relation between the spectral distortion and the MVDR model order is one issue which needs to be understood and could potentially lead to further performance improvements. Another potential application of this approach currently under investigation is the discrimination of voiced and unvoiced speech.

7. REFERENCES

[1] P. Stoica and R. Moses, *Spectral analysis of signals*, Prentice Hall, NJ, USA, 2005.

[2] D. G. Monolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing*, McGraw-Hill, USA, 2000.

[3] S. Dharanipragada and B. D. Rao, "MVDR based feature extraction for robust speech recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Utah, May 2001, vol. 1, pp. 309–312.

[4] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8(3), pp. 221–239, May 2000.

[5] P. J. Sherman and K. N. Lou, "On the family of ML spectral estimates for mixed spectrum identification," *IEEE Trans. Signal Processing*, vol. 39(3), pp. 644–655, March 1991.

[6] NTIS, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.

[7] R.J McAulay, T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 744–754, Aug 1986.