MINIMUM VARIANCE DISTORTIONLESS RESPONSE (MVDR) MODELING OF VOICED SPEECH

Manohar N. Murthi *Bhaskar D. Rao*

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

ABSTRACT

In this paper we propose the MVDR method, which is based upon the Minimum Variance Distortionless Response (MVDR) spectrum estimation method, for modeling voiced speech. Developed to overcome some of the shortcomings of Linear Prediction models, the MVDR method provides better models for medium and high pitch voiced speech. The MVDR model is an all-pole model whose spectrum is easily obtained from a modest non-iterative computation involving the Linear Prediction coefficients thereby retaining some of the computational attractiveness of LPC methods. With the proper choice of filter order, which is dependent on the number of harmonics, the MVDR spectrum models the formants and spectral powers of voiced speech exactly. An efficient reduced model order MVDR method is developed to further enhance its applicability. An extension of the reduced order MVDR method for recovering the correct amplitudes of the harmonics of voiced speech is also presented.

1. INTRODUCTION

In speech compression systems, good modeling of the short-term spectrum of speech is critical for success. For voiced speech, which is periodic, a speech compression system must model the powers of the spectrum at the harmonic frequencies well, especially the locations and powers of the perceptually important formants. For both time-domain and frequency-domain coders, the Linear Prediction (LP) method has been used to model the short-term spectrum of speech [1], [2]. The LP method has achieved remarkable success and longevity in speech compression systems because of its simplicity of computation, ability to provide a spectral envelope for both voiced and unvoiced speech, and its amenability to efficient Vector Quantization techniques. However, the LP spectrum does not model voiced speech well, especially medium and high pitch voiced speech [3], [4]. In particular, the LP spectrum does not model the spectral powers at the harmonic frequencies well, especially at the perceptually important formants.

To address these shortcomings, we introduce the MVDR method as an alternative and complement to the LP method. The MVDR method is based upon the Minimum Variance Distortionless Response Spectrum, a spectrum estimation method popular in array processing [5]. In contrast to the LP filter, the MVDR method, using a high order all-pole filter, provides a spectrum that models exactly the spectral powers at the harmonic frequencies of voiced speech. In addition, the MVDR method is a computationally simple technique based upon the LP coefficients themselves. As a result, speech compression systems can employ the MVDR method without abandoning the popular methods for quantizing LP coefficients. In addition, the MVDR method may be used to parametrically represent voiced speech spectral powers in a frequency domain speech coder such as the Multi-Band Excitation coder [6].

This paper is organized as follows. In section II, we consider the limitations of Linear Prediction in modeling voiced speech. In section III, we present the MVDR method and discuss its suitability for modeling voiced speech. In section IV, we address some practical considerations in MVDR modeling, namely model order reduction, and refinement of the power estimates of voiced speech harmonics.

2. THE LINEAR PREDICTION METHOD

An all-pole filter is commonly employed to model the vocal tract in speech processing. The Linear Prediction method is used to obtain the parameters of the all-pole filter. The LP filter tries to whiten the input signal by minimizing the variance of its output. The LP filter is defined as $A(z) = 1+\sum_{k=1}^{M} a_k z^{-k}$ where the a_k's are LP coefficients obtained by the Levinson-Durbin computation, and M is the filter order. In addition, the LP filter defines a spectrum,

$$P_{LP}(\omega) = \frac{P_e}{\left|1+\sum_{k=1}^{M} a_k e^{-j\omega k}\right|^2} \tag{1}$$

where P_e is the prediction error variance. We consider analytically the performance of the LP method in modeling voiced speech by examining its performance for a signal with a discrete line spectrum, which is a reasonable approximation of a voiced speech signal.

In particular, assume that the voiced speech signal is periodic, and for short times, usually 20ms, is modeled as

$$u(n) = \sum_{k=1}^{L} c_k cos(\omega_0 kn + \phi_k) \tag{2}$$

where ω_0 is the fundamental (radial) frequency of the speech, the c_k's are the amplitudes at the harmonic frequencies, and L is the number of harmonics. The pitch of the speech is $f_0 = \frac{\omega_0}{2\pi}$, and the number of harmonics

is $L = \lfloor \frac{f_s}{2f_0} \rfloor$ where f_s is the sampling frequency, typically 8kHz. The number of harmonics L decreases as the pitch increases. With this model for voiced speech, the signal has a correlation sequence

$$r_{uu}(m) = \sum_{k=1}^{L} \frac{|c_k|^2}{4} cos(\omega_0 k m). \quad (3)$$

This model for voiced speech exhibits a discrete line spectrum with the spectral powers $S_{uu}(\omega_0 k) = \frac{|c_k|^2}{4}$ at the harmonic frequencies, $\omega_0 k$, $1 \leq |k| \leq L$.

For this voiced speech model, the prediction error, which is the variance of the output of the LP filter $A(z)$, is given by

$$P_e = \sum_{k=1}^{L} \frac{|c_k|^2}{2} |A(e^{j\omega_0 k})|^2, \quad (4)$$

where we have used the fact that the LP filter is real and has a symmetric frequency response. P_e can be minimized by placing the zeros of $A(e^{j\omega})$ at $e^{j\omega_0 l}$, which corresponds to the harmonic content of the signal. P_e can be made zero for a signal with L harmonics, and consequently $2L$ exponentials, when the filter has enough zeros to cancel all the input exponentials, i.e. $M \geq 2L$. Therefore, for $M \geq 2L$ the LP spectrum (c.f. eq. 1) or even $\frac{1}{|A(e^{j\omega})|^2}$ does not give any indication of the power at the harmonics. In general, for M near $2L$, which corresponds to the high to medium pitch voiced speech case, one can foresee modeling problems using the LP method.

In the case where the filter order is not sufficient, or $M < 2L$, the LP filter does not have enough filter zeros to cancel the $2L$ exponentials in the input signal. For $M < 2L$, the following relationship can be established between the LP spectral estimate at a particular harmonic frequency $\omega_0 l$ and the power of the harmonics by substituting the expression for the prediction error (c.f. eq. 4) into the expression for the LP spectrum (c.f. eq. 1),

$$P_{LP}(\omega_0 l) = \frac{|c_l|^2}{2} + \sum_{k=1, k\neq l}^{L} \frac{|c_k|^2}{2} \frac{|A(e^{j\omega_0 k})|^2}{|A(e^{j\omega_0 l})|^2}. \quad (5)$$

By some simple manipulations, it can be shown that

$$\sum_{k=1}^{L} \frac{\frac{|c_k|^2}{2}}{P_{LP}(\omega_0 k)} = 1. \quad (6)$$

This sets up a complicated relationship between the LP spectral estimates $P_{LP}(\omega_0 k)$, and the actual spectral powers, $\frac{|c_k|^2}{4}$. The LP filter does not have a constrained response at any particular frequency, and so the prediction error and consequently the spectral estimate can be somewhat arbitrary at the harmonic frequencies. This relationship creates problems in attempts to compensate the LP spectral estimates at the harmonic frequencies. For $M << 2L$, an alternative argument can be used to show that the LP method is a reasonable approach and that the LP spectrum scaled by $\frac{1}{2L}$ does yield a reasonable estimate of the powers at the harmonics. Consequently for low pitch signals the LP spectrum better approximates the original voiced speech spectrum, and the biasing is less severe.

To see some of the problems of LP in modeling voiced speech, consider Figure 1. We used real voiced speech signals in all of our simulations. In addition, we used autocorrelation sequences like the one in eq. 3 by employing a simple peak-picking method and open-loop pitch estimation. In this example, a 14th order LP spectrum is modeling a 320Hz voiced speech spectrum. The LP spectrum overestimates the spectral power at one of the main formants. A higher order LP filter will not correct the overestimation, and in fact will only exacerbate the problem.

3. MVDR MODELING OF VOICED SPEECH

The MVDR method is similar to the LP method in many respects in that the MVDR filter is also an all-pole filter that is obtained by an extension of the Levinson-Durbin computation [5]. In contrast to the LP method which constrains the first coefficient of its filter to be 1, in the MVDR method an FIR filter $h(n)$ is designed which minimizes its output variance subject to the constraint that the response of the filter at a particular frequency ω_l has unity gain, namely $H(e^{j\omega_l}) = \sum_{n=0}^{M} h(n) e^{-j\omega_l n} = 1$. The constraint is commonly referred to as the distortionless constraint. Mathematically, the optimum FIR filter, denoted by $h_l(n)$, for frequency ω_l is obtained as a solution to the optimization problem

$$\min_{h(n)} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 S_{uu}(e^{j\omega}) d\omega \right) \text{ subj. to } H(e^{j\omega_l}) = 1.$$

$S_{uu}(e^{j\omega})$ is the actual power spectrum of the input signal. The power at the output of the optimized constrained filter is used by the MVDR method as an estimate of the power spectrum at the frequency ω_l, i.e.

$$P_{MV}(\omega_l) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_l(e^{j\omega})|^2 S_{uu}(e^{j\omega}) d\omega. \quad (7)$$

The distortionless response constraint at ω_l coupled with the output power minimization ensures that the MVDR filter $h_l(n)$ designed for ω_l will attempt to faithfully preserve the input signal power at ω_l. This is the key difference between the MVDR method and the LP method that allows the MVDR method to give a better model for high pitch voiced speech.

We must stress that in the MVDR method the need to design a separate filter for each frequency, ω_l, is only conceptual. It can be shown that the MVDR spectrum for all frequencies can be simply computed as

$$P_{MV}(\omega) = \frac{1}{\mathbf{v}^H(\omega) \mathbf{R}_{M+1}^{-1} \mathbf{v}(\omega)} = \frac{1}{|B(e^{j\omega})|^2}, \quad (8)$$

where $\mathbf{v}(\omega) = [1, e^{-j\omega}, e^{-j2\omega}, \cdots, e^{-jM\omega}]^T$, and $\mathbf{R}_{M+1}$ is the $(M+1) \times (M+1)$ Toeplitz autocorrelation matrix of the input signal [5]. This spectrum is easily obtained from a modest non-iterative computation using the LP coefficients [5]. So a speech compression system can still use the Levinson-Durbin computation to obtain the LP coefficients, and can still use the LP Vector Quantization methods. The

all-pole filter $B(z)$ corresponding to the MVDR spectrum is also easily obtainable using a non-iterative spectral factorization technique that does not involve root-searches [7].

We now elaborate on the MVDR method and voiced speech modeling. For any voiced signal of pitch ω_0, from eq. 7, the MVDR spectral estimate at $\omega_l = \omega_0 l$ is given by

$$P_{MV}(\omega_0 l) = \sum_{k=1}^{L} \{|H_l(e^{j\omega_k})|^2 + |H_l(e^{-j\omega_k})|^2\} \frac{|c_k|^2}{4},$$

where L is the number of harmonics. The MVDR filter $h_l(n)$ designed for the lth harmonic $\omega_0 l$ will try to faithfully preserve the input power at $\omega_0 l$ while trying to mitigate the influence of the other $2L-1$ exponentials. In particular,

$$P_{MV}(\omega_0 l) = \frac{|c_l|^2}{4} + |H_l(e^{-j\omega_l})|^2 \frac{|c_l|^2}{4} + \sum_{k=1, k\neq l}^{L} \{|H_l(e^{j\omega_k})|^2 + |H_l(e^{-j\omega_k})|^2\} \frac{|c_0 k|^2}{4}. \quad (9)$$

The first term is the power in the exponential at $\omega_0 l$ which passes through the filter undistorted because of the distortionless constraint. If the MVDR filter has M filter zeros, and $M \geq 2L-1$, then the MVDR filter has enough degrees of freedom to cancel all the other input exponentials. In this case, the MVDR spectral estimate obtains the exact value of the input spectral power at that harmonic. When the filter order M is insufficient, or $M < 2L-1$, the MVDR filter does not have enough zeros to cancel out the interfering harmonic signals. In this case, the MVDR spectral estimate will exhibit a positive bias as the other exponential signals leak through.

Based on the above discussion, the power estimates at the harmonics get better as the model order M increases. Consequently, in modeling voiced speech, the MVDR filter with a sufficient order usually models voiced speech better than the LP filter. In particular, we have observed that for most cases of voiced speech, when the MVDR filter order M is greater than the number of harmonics L, the MVDR spectrum outperforms the LP spectrum in modeling the formants and spectral powers at the harmonics. In addition, we have noticed excellent results for a filter order of about $M = 1.6L$. For instance, consider Figure 2. In this case, the MVDR spectrum using a filter order of $M = 41$ models the 160Hz medium pitch voiced speech spectrum very well. Note that the MVDR spectrum captures the spectral powers at the main formants exactly, and models most of the spectral powers at the harmonic frequencies quite well. The MVDR spectrum exhibits a modest bias in the less perceptually important valleys of the spectrum. As the filter order is increased, the MVDR spectrum improves, better matching the input speech spectrum. This is consistent with good modeling behavior because it is important that models perform better as the model order is increased. In contrast, the LP spectrum does not have such a feature.

For high pitch speech, the MVDR method does not need a large filter order, and so it performs better than the LP method. For low pitch, the LP approach is appropriate. For medium range pitch, the crossover point is unclear and needs further study. Now we consider modifications that enhance the features of MVDR modeling.

4. MODEL ORDER REDUCTION AND AMPLITUDE RECOVERY

We now discuss an approach to reducing the model order and an approach to recovering the harmonic powers.

4.1. Reduced Order MVDR Filters

In the MVDR method, we have seen that the high order filter is able to capture the spectral envelope and spectral powers at the harmonics quite well. It is desirable to translate the quality of the high order estimate to a low order filter. The high order MVDR filter models an AR spectrum $P_{MV}(\omega)$ that provides a good spectral envelope for medium and high pitch voiced speech. We note that $P_{MV}(\omega)$ is strictly speaking not a power spectrum in that its Fourier inverse does not give back the input autocorrelation sequence. Consequently, approximating the higher order MVDR spectrum by a lower order filter does not simply amount to applying the MVDR method to a small set of given autocorrelation lags. Instead, we can view the high order MVDR spectrum as a very good AR model of the vocal tract and it is this high quality MVDR spectrum that needs to be approximated by a low order all-pole filter.

The order reduction procedure is as follows. First, we take a high order, high quality MVDR spectrum, and obtain an all-pole, causal and stable filter $B(z)$ from it by using a non-iterative spectral factorization [7]. The Levinson-Durbin procedure is run in reverse to take the high order $B(z)$, and obtain a lower-order filter $C(z)$ that approximates $B(z)$. This reduced order MVDR filter $C(z)$ often outperforms the normal low order LP filter in modeling medium to high pitch voiced speech. Rather than trying to model a discrete line spectrum, the reduced order MVDR spectrum will try to match the original high quality MVDR spectrum.

For example, consider Figure 3. In this case, the 14th order reduced order MVDR spectrum based upon the high order ($M = 31$) MVDR spectrum models the 186Hz voiced speech spectrum better than a normal 14th order LP spectrum, especially at the formant frequencies. If we raise the order of the reduced order MVDR spectrum, its estimates become better. In fact, the performance of the reduced order MVDR spectrum depends directly upon the order mismatch between the reduced order MVDR filter, and the high order high quality MVDR filter.

4.2. Amplitude Recovery in the MVDR Method

In frequency-domain coders such as the Multi-Band Excitation coder, better estimates of the voiced speech spectral powers at the harmonic frequencies are desired. We can use some properties of the MVDR approach to obtain refined estimates from the reduced order MVDR models.

Recall that if the input voiced speech is assumed to be a discrete line spectrum, then the MVDR spectral estimate at harmonic frequency $\omega_l = \omega_0 l$ is given by eq. 9. We can obtain estimates for all L harmonic frequencies, $\omega_0, \cdots, \omega_0 L$ in which the estimates $P_{MV}(\omega_0 l)$ and filter responses $H(e^{j\omega_0 l})$ are known and where the true spectral powers $\frac{|c_k|^2}{4}$, $k = 1, .., L$ are the unknowns. Using eq. 9,

one can set up a system of L equations with L unknowns. This can be written compactly as $\mathbf{p} = \mathbf{Wc}$, where $\mathbf{p}$ is the Lx1 column vector of MVDR spectral estimates, $P_{MV}(\omega_0 l)$, $l = 1, .., L$, $\mathbf{W}$ is the matrix of MVDR filter gains, and $\mathbf{c}$ consists of the unknown actual spectral powers. We are interested in the case where the filter order M is less than the number of harmonics L. In this case, the rank of the LxL matrix $\mathbf{W}$ is just $M + 1$. Hence $\mathbf{W}$ is not invertible. However, we can obtain a minimum norm solution by setting $\hat{\mathbf{c}} = \mathbf{W}^{\dagger}\mathbf{p}$ where $\dagger$ denotes the pseudo-inverse of the matrix. In many cases, this minimum norm solution can recover some of the true spectral powers at the harmonics for medium to high pitch speech, and can provide a better estimate than a comparable order LP filter.

For example, consider Figure 4. In this case, the 14th order compensated MVDR spectral estimates of the 222 Hz voiced speech powers at the harmonic frequencies are better than the normal LP estimates. Hence, the original spectral powers at the harmonic frequencies can be better recovered from a reduced order MVDR spectral estimate by using this matrix compensation procedure.

REFERENCES

[1] Markel, J.D. and A.H. Gray, *Linear Prediction of Speech*, Springer Verlag, 1976.

[2] J. Makhoul, "Linear Prediction: A Tutorial View," *Proceedings of the IEEE*, April 1975.

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

[4] D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley, 1987.

[5] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, 1996.

[6] A.M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, Wiley, 1994.

[7] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1993.

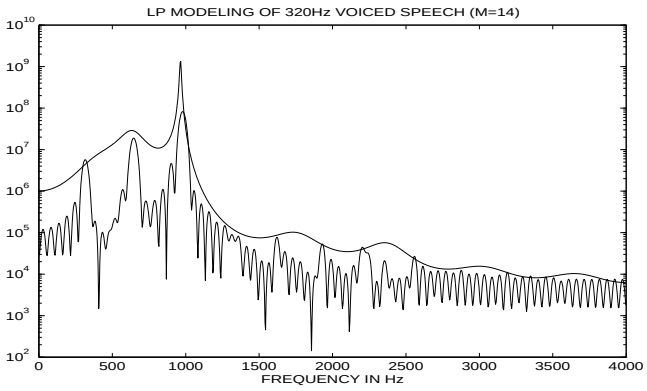

Figure 1. Linear Prediction Spectrum of a 320Hz Voiced Speech Spectrum, M=14

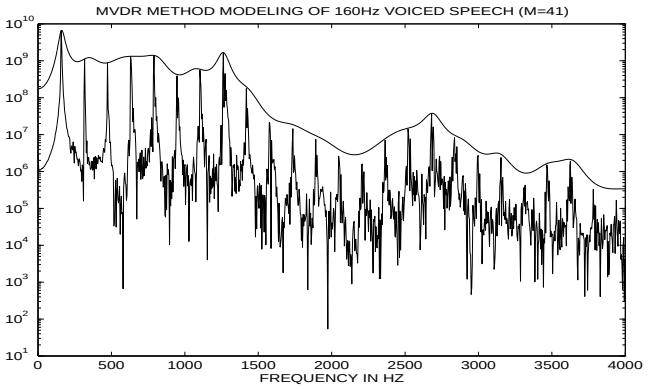

Figure 2. MVDR Spectrum of a 160Hz Voiced Speech Spectrum, M=41

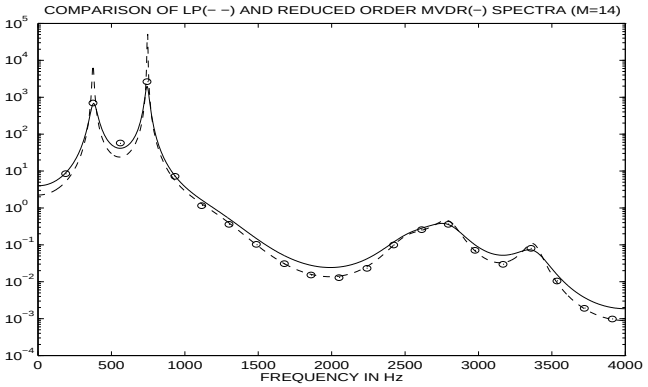

Figure 3. Comparison of normal LP spectrum (- -), reduced order MVDR spectrum, and 186Hz Voiced Speech Harmonics (o), M=14

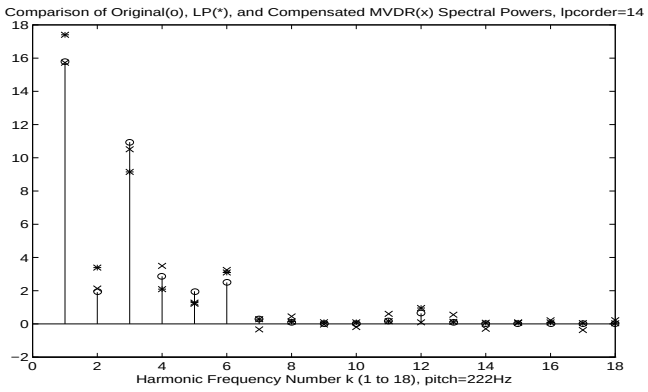

Figure 4. Comparison of original 222Hz voiced speech(o), LP(*), and compensated MVDR(x) spectral powers at the harmonics. The Number of harmonics L=18, and M=14