TOWARDS A SYNERGISTIC MULTISTAGE SPEECH CODER

Manohar N. Murthi and Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego
manoharn@ece.ucsd.edu brao@ece.ucsd.edu http://raman.ucsd.edu

ABSTRACT

In this paper, we propose some new modeling techniques that provide a more synergistic approach to multistage time-domain speech compression. In particular, we propose a new error criterion for determining all-pole filters, and a unique method for jointly coding the pulse information in excitation vectors. The new error criterion for determining all-pole filters is based upon minimizing the sum of the residual signal's absolute values raised to a power less than one. It is shown to be a desirable cost function for yielding residual signals that are more sparse, and consequently better suited for multistage compression than Linear Prediction residuals. Statistical reasons supporting the new criterion are also provided. Furthermore, exploiting the properties of, and the relationship between, the Linear Prediction and Minimum Variance spectra, we propose a novel parameter set for jointly coding the excitation vector's pulse position, sign, and gain information.

1. INTRODUCTION

In speech compression, the class of Linear Prediction Analysis by Synthesis (LPAS) coders has achieved much success [1]. In LPAS coders, the task of compression is broken into several stages. We propose some new modeling techniques that attempt to provide tighter coupling between the multiple stages of time-domain speech coders, and thereby provide a framework for moving towards truly synergistic multistage coders that enable more efficient coding.

In particular, we propose a new error criterion for determining all-pole filters, and a unique method for jointly coding the pulse information in codebook excitation vectors. The new error criterion for determining all-pole filters, $1/A(z)$, constructs corresponding first-stage analysis filters, $A(z)$, that explicitly try to produce sparse residual signals with few dominant non-zero values. In particular, the filter tries to minimize the sum of the residual signal's absolute values raised to a power p less than one, i.e. minimize $\sum |r_i|^p,\ 0 < p < 1$. In contrast, existing LPAS coders use Linear Prediction analysis filters corresponding to a value of $p = 2$ which yield residuals that are not explicitly well-suited for encoding by the pitch and codebook stages. Both deterministic cost function and statistical arguments are shown to justify the new error criterion, and its promotion of sparse residuals.

With sparse residuals from an FIR analysis filter, a parsimonious representation of the codebook excitation is desirable. We propose a novel method for jointly coding pulse position, sign, and gain information of excitation vectors in time-domain coders. This method is based on an exploitation of the properties of, and the relationship between, the Linear Prediction and Minimum Variance Distortionless Response (MVDR) Spectra [7]. In most LPAS coders, the codebook excitation pulse positions, signs, and amplitudes are quantized separately. The technique proposed for jointly modeling the excitation vector information allows for systematic tradeoffs between bit allocation and codebook excitation accuracy in a manner hitherto not possible.

This research was supported by UC MICRO Grant 97-146 and by Hughes Electronics.

2. ALL-POLE FILTERS WITH SPARSE RESIDUALS

In this section, we present a new method for obtaining FIR analysis filters which produce sparse residuals with few large energy entries. First, we consider the popular speech production model. In most time-domain coders, a frame of speech $s(n), 0 \leq n \leq N-1$, is modeled as an Auto-Regressive (AR) process,

$$s(n) = \sum_{k=1}^{M} \alpha_k s(n-k) + e(n) \tag{1}$$

where the α_k's are the AR parameters, and $e(n)$ is the driving noise process.

The speech production model (Eq. 1) in matrix form is

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{r}, \tag{2}$$

where $\mathbf{y}$ is the $N \times 1$ vector containing the N speech samples $s(n)$, $\mathbf{H}$ is the $N \times M$ matrix of speech samples in which the ith column contains the samples ranging from $s(-i), \cdots, s(N-(i+1))$, $\mathbf{x}$ is the $M \times 1$ vector of AR parameters α_k, and $\mathbf{r}$ is the $N \times 1$ residual, consisting of the excitation sequence, $e(n)$.

The filter parameters $\mathbf{x} = [\alpha_1 \alpha_2 \cdots \alpha_M]^T$, corresponding to an analysis filter of $A(z) = 1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} \cdots - \alpha_M z^{-M}$, are found by minimizing a function of the residual $\mathbf{r}$.

2.1. A New Error Criterion For Sparse Residuals

To produce sparse residuals, the filter parameters of $A(z)$ are determined by the new error criterion

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \lim_{p \to 0} \sum_{i=1}^{N} |r_i|^p, \tag{3}$$

where $r_i,\ 1 \leq i \leq N$ are the elements of $\mathbf{r}$. As p goes towards zero, the influence of the small values in the residual increases while the influence of the large residual values decreases, and the sum counts the number of non-zero elements, the true measure of

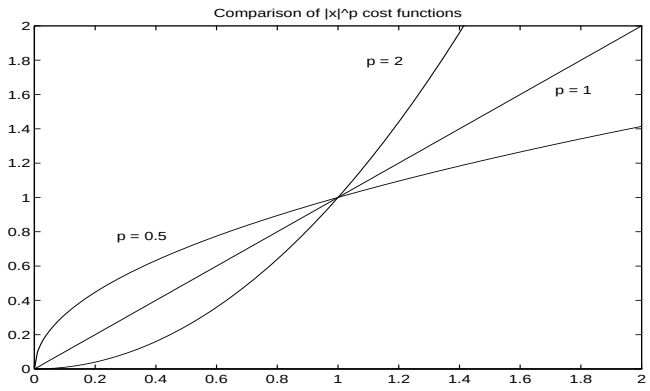

Figure 1: Comparison of $|x|^p$ cost functions

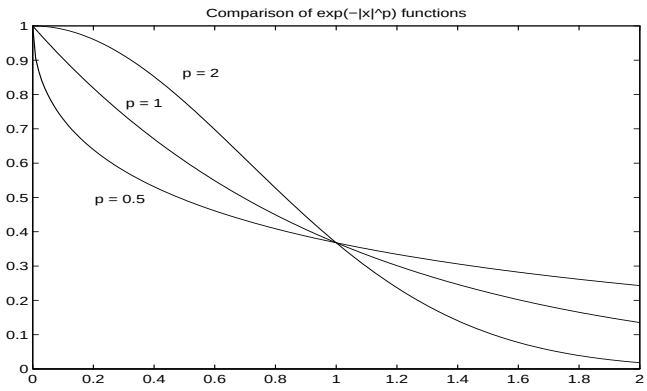

Figure 2: Comparison of $e^{-|x|^p}$ functions

sparsity. The cost function provides an incentive for $\hat{\mathbf{x}}$ to produce residuals with values close to zero.

In Linear Prediction, a value of $p = 2$ is fixed in Eq. 3, and corresponds to the minimization of the ℓ_2 norm of $\mathbf{r}$. By minimizing the sum of the squares of the residual, $\mathbf{r}$, the Linear Prediction filter is very sensitive to, and over-influenced by, large errors which are common for voiced speech pitch pulses, thereby producing residuals that are not sparse in nature and are not ideal for compression purposes.

Previous attempts to change the error criterion employed in all-pole filter parameter estimation for time-domain coding have centered around minimizing Eq. 3 for fixed values of $p, 1 \leq p \leq 2$ [2], [3], although in [4], the sum of a weighted residual was minimized. Minimizing the $\ell_p, 1 \leq p \leq 2$ norms provides filters which dampen the influence of the large residuals, but not to a large degree. For example consider Figure 1. In this figure, the cost functions for three values of p are compared. It is clear that a value of $p < 1$ provides more incentives for providing small residuals, because the small residuals are weighted more heavily. In contrast, the conventional value of $p = 2$ which corresponds to the Linear Prediction cost function weights large residual values more than small residual values, thereby providing overall residual signals that are not sparse in nature.

In practice, a value of p must be fixed in Eq. 3, and in general, we explore $0 < p < 1$. Although the resulting sparsity measure is not strictly a norm, it does yield analysis filters with sparse residuals. We now provide a statistical justification for choosing $0 < p < 1$.

2.2. Statistical Interpretation of the Sparse Residual Measure

The Linear Prediction solution can be viewed as the Maximum Likelihood Estimate (MLE) of an AR process driven by a Gaussian noise sequence $\mathbf{r}$. The minimization of the ℓ_1 norm of $\mathbf{r}$ can be viewed as the Maximum Likelihood Estimate of an AR process driven by Laplacian noise. For $\ell_p, 1 \leq p \leq 2$ minimization, the sequence $\mathbf{r}$ driving the AR process can be viewed as having a generalized Gaussian density function $ce^{-|x|^p}$, of which the Laplacian is a special case. With values of $1 \leq p \leq 2$, density functions of the residual sequence have heavier tails than the Gaussian distribution, and consequently admit larger errors in the residual. However, in previous studies, values of $p < 1$ were not considered.

From a statistical point of view, values of $p < 1$ are desirable because the density functions in these cases have a sharper slope near zero, and have heavier tails. A peak at zero and a steep slope at zero indicate that the likelihood function is quite sensitive to values near zero and its maximization should encourage small values to become smaller. On the other hand, the heavy tails along with the lowered sensitivity of the likelihood to large values should support a few large entries. For example, consider Figure 2. In this figure, the function $e^{-|x|^p}$ is plotted for values of $p = 0.5, 1, 2$. It is apparent that for the value of $p = 0.5$, the function has a much sharper slope at values of x close to zero, and has heavier tails than the other functions. Consequently, by allowing $0 < p < 1$, the new error criterion is geared towards sparse residuals from a likelihood viewpoint.

2.3. Computation with the Sparse Residual Measure

For a given value of $0 < p < 1$, Eq. 3 must be computed. There is no closed form solution and therefore numerical methods must be employed. The solution can be found using the Iteratively Reweighted Least Squares (IRLS) algorithm [5], in which a sequence of solutions to weighted least squares problems must be computed. In particular, the solution $\hat{\mathbf{x}}$ is found by

$$\begin{aligned} \text{for } k = &\quad 0, 1, 2, \cdots \\ \mathbf{r}^{(k)} = &\quad \mathbf{y} - \mathbf{H}\mathbf{x}^{(k)} \\ \mathbf{D_k} = &\quad \text{diag}((|\mathbf{r}^{(k)}|)^{(p-2)/2}) \\ \mathbf{x}^{(k+1)} = &\quad \arg\min_{\mathbf{x}} \|\mathbf{D_k}(\mathbf{y} - \mathbf{H}\mathbf{x})\|_2 \\ k = &\quad k + 1 \end{aligned}$$

where $\mathbf{x}^{(0)}$ is initialized to the Linear Prediction solution. It can be shown that $\sum_{i=1}^{N} |r_i^{(k+1)}|^p \leq \sum_{i=1}^{N} |r_i^{(k)}|^p$ even when $p < 1$, meaning that this is a descent algorithm.

2.4. Simulation Results

In practice, the utilization of the sparse residual measure in all-pole filter design leads to analysis filters whose residuals often feature sharper pitch spikes and more sparsity than corresponding Linear Prediction filter residuals, especially for voiced speech. For example, consider Figure 3. Here the original speech is shown on top, the 30th order Linear Prediction residual is shown in the

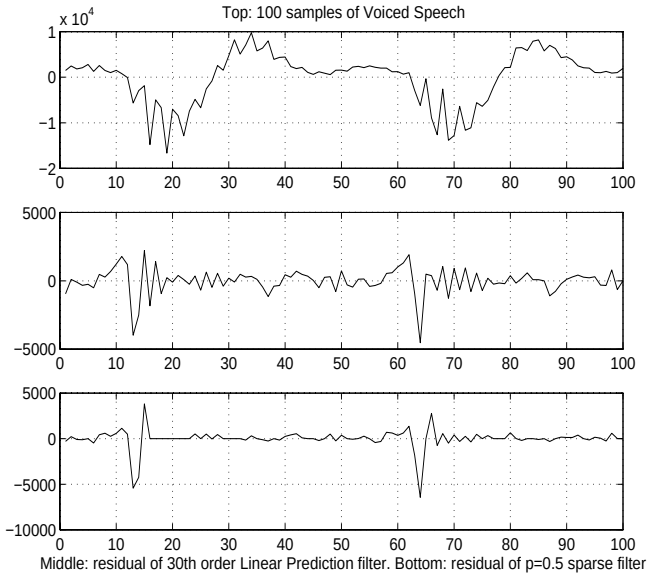

Figure 3: Original voiced speech (top), 30th order Linear Prediction residual (middle), new filter (30th order) with sparse residual measure, p=0.5 (bottom)

middle, and the new filter's residual signal with $p = 0.5$ is shown at the bottom.

In this example, the benefits of the new filter are quite apparent as its residual has most of its energy concentrated in a few places corresponding to multiples of the pitch. In general, for a large filter M to frame size N ratio, the new error criterion based on a sparse measure does provide a filter that yields sparse residuals, especially for voiced speech. This is consistent with the goals of multistage time-domain coders in which a sparse residual is desired. For implementation and utilization in a practical compression system, more study is needed as the filter order and interpolative properties become more important.

3. A NOVEL METHOD FOR MODELING EXCITATION INFORMATION

With a sparse residual, the task of pitch and codebook excitation determination is made simpler. The codebook excitation can be determined by using basis selection methods to find a sparse solution to a linear system [6]. Now we present a unique method for jointly coding the codebook excitation information in time-domain speech coders that achieves a parsimonious parametric representation. Typical excitation vectors in Multi-Pulse and Algebraic CELP coders contain a few pulses each with associated gain and sign information. For example, for a subframe of size $N_s = 40$, usually $K = 5$ pulses are employed, each pulse with its own position, sign, and for Multi-Pulse coders, each with its own gain. In most coders, the pulse positions, signs, and gains are quantized separately.

Using properties of, and the relationship between the Linear Prediction filter and the Minimum Variance Distortionless Response (MVDR) spectrum, [7], we propose a novel method for jointly coding the K pulse positions, signs, and gains. We show that the K pulse positions, K signs, and K gains can be exactly represented by $2K-1$ reflection coefficients and one prediction error variance that both correspond to a single filter. We now discuss some properties of Linear Prediction and MVDR spectra relevant to the encoding scheme.

3.1. Properties of Linear Prediction and MVDR Spectra

First we define an input signal consisting of the sum of K real cosine signals, $u(n) = \sum_{i=1}^{K} c_i \cos(\omega_i n)$ and corresponding correlation sequence $r_{uu}(m) = \sum_{i=1}^{K} 2S(\omega_i)\cos(\omega_i m)$, where $S(\omega_i) = |c_i|^2/4$. The input signal exhibits a discrete line spectrum at the positive and negative frequencies $\pm\omega_i, i = 1, \cdots, K$ with spectral powers $S(\omega_i)$.

Lemma 1. Define $r(m) = \sum_{i=1}^{K} 2S(\omega_i)\cos(\omega_i m)$ a correlation sequence, a Linear Prediction filter $A_M(z)$ with order $M = 2K$ has zeros at the $2K$ positive and negative frequencies $\pm\omega_i, i = 1, \cdots, K$, i.e. $A_M(e^{\pm j\omega_i}) = 0$.

Therefore, an Mth order Linear Prediction filter places its filter zeros at the frequencies of the line spectrum, providing accurate frequency location information, but not amplitude information since for any of the spectral line frequencies, $A_M(e^{\pm j\omega_i}) = 0$.

Now we state a property of the MVDR spectrum that we exploit for modeling amplitude information. The MVDR spectrum is also known as the Minimum Variance spectrum, or Capon's Method.

Lemma 2. Define $r(m) = \sum_{i=1}^{K} 2S(\omega_i)\cos(\omega_i m)$, a correlation sequence. The MVDR spectrum $P_{MV}^{(M)}(\omega)$ of order $M = 2K-1$ models the powers of the line spectra exactly, i.e. $P_{MV}^{(2K-1)}(\omega_i) = S(\omega_i)$.

Details of MVDR spectral modeling of exponentials are in [8]. With the results on Linear Prediction modeling of frequency location information, and MVDR modeling of amplitude information, we can state the following.

Theorem. Define $r(m) = \sum_{i=1}^{K} 2S(\omega_i)\cos(\omega_i n)$ with $\omega_i \neq 0, \pi$. Then the Prediction error variance $P_e^{(2K-1)}$ and reflection coefficients $\Gamma_m, m = 1, \cdots, 2K-1$ corresponding to a $2K-1$ order Linear Prediction filter $A_{2K-1}(z)$ based on $r(m)$, are sufficient to recover the line frequency locations ω_i and the spectral powers $S(\omega_i)$ exactly.

Outline of Proof. Note that from the given $r(m)$ and the constraints $\omega_i \neq 0, \pi$, we know that $\Gamma_{2K} = 1$. Consequently the given $\Gamma_m, m = 1, \cdots, 2K-1$, are sufficient to construct the order $2K$ Linear Prediction filter $A_{2K}(z)$. From Lemma 1, $A_{2K}(z)$ has its filter zeros at the frequencies ω_i. The $\Gamma_m, m = 1, \cdots, 2K-1$ and $P_e^{(2K-1)}$ can be used to obtain the order $(2K-1)$ MVDR spectrum [7]. From Lemma 2, $P_{MV}^{(2K-1)}(\omega)$ models the spectral powers exactly at the line frequencies, ω_i.

3.2. Joint Pulse Position and Amplitude Coding

Consider the coding of K pulses from an excitation vector of length N_s. Typical values are $K = 5$ and $N_s = 40$. The K pulses have positions $p_1, p_2, \cdots, p_K$, and are allowed to range from $1 \leq p_i \leq N_s$. Each pulse p_i is weighted by $s_i g_i$ where s_i is the sign of the pulse, and g_i is its positive gain value.

Part I: Encoding.

For ease of presentation, we first consider a case where all the signs are positive, i.e. $s_i = 1$.

Step 1. An autocorrelation sequence is constructed as follows

$$r(m) = \sum_{i=1}^{K} 2g_i \cos(\pi \frac{p_i}{N_s + 1} m). \quad (4)$$

This sequence has a corresponding discrete line spectrum in the frequency domain. The "sampling frequency" corresponds to $N_s + 1$ and the spectral peaks are positioned at frequencies $\omega_i = \pi p_i / (N_s + 1)$ which encode the pulse position information. The line spectra at the frequencies ω_i have corresponding powers g_i which encode the amplitude information.

Step 2: The $2K - 1$ order Linear Prediction filter reflection coefficients $\Gamma_m, m = 1, \cdots, 2K - 1$ and corresponding prediction error variance $P_e^{(2K-1)}$ are computed using $r(m)$ from Eq. 4, and sent to the decoder.

Step 2 is motivated by the above theorem which suggests that the parameter set computed is adequate for recovering pulse position and amplitude information.

Part II: Decoding.

Step 1: The order $2K$ Linear Prediction filter $A_{2K}(z)$ is constructed from the given $\Gamma_m, m = 1, \cdots 2K-1$, and with $\Gamma_{2K} = 1$. From Lemma 1, the zeros of $A_{2K}(z)$ are found at the frequencies ω_i which give pulse position information.

Step 2: The order $2K - 1$ MVDR spectrum is directly computed from the given Γ_m values and prediction error variance $P_e^{(2K-1)}$ [7]. From Lemma 2, the $2K - 1$ order MVDR spectrum models the gain information exactly, i.e. $P_{MV}^{(2K-1)}(\omega_i) = g_i$.

We can easily incorporate both positive and negative signs by modifying the construction of the correlation sequence. By utilizing multiples of $0.5\pi/(N_s + 1)$, we can incorporate sign information. For negative signs, we follow the convention of adding the fractional frequency value $0.5\pi/(N_s+1)$ to the original line frequency. In particular, if $s_i = -1$, we add $0.5\pi/(N_s+1)$ to the original frequency $\pi p_i/(N_s+1)$ to obtain a new line frequency value $\pi(p_i + 0.5)/(N_s + 1)$ that is used in the construction of the correlation sequence in Eq. 4.

3.3. Example of Joint Coding

Consider encoding $K = 5$ pulses at positions $(5, 7, 8, 12, 30)$ with corresponding gains $(-1, 10, -7, 4, 3)$. The correlation sequence in Eq. 4 with the sign modifications is constructed and the $(2K-1) = 9$th order Linear Prediction reflection coefficients and $P_e^{(2K-1)}$ are transmitted by the encoder. The decoder computes the $2K = 10$th order Linear Prediction filter from the transmitted reflection coefficients and prediction error variance, using $\Gamma_{2K} = 1$.

The Linear Prediction filter's frequency response is evaluated at $2(N_s + 1) = 82$ points corresponding to frequencies $\omega_k = \pi k/(N_s + 1), k = 0, 0.5, 1, 1.5, 2, \cdots, 40, 40.5$ as shown in the top of Figure 4. The bins in the figure are numbered from 1 to 82. Consequently 5 zeros of the LP analysis filter are discovered at bins $(12, 15, 18, 25, 61)$. This corresponds to positions $(5.5, 7, 8.5, 12, 30)$. The presence of non-integer values indicates

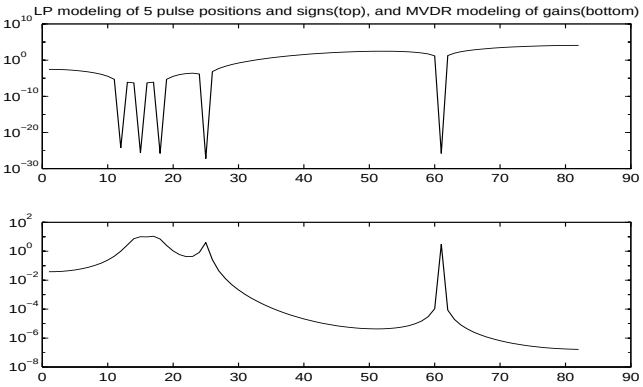

Figure 4: At the top, The Linear Prediction analysis filter is evaluated at $2(N_s + 1)$ points in the frequency domain. The filter zeros are used to determine the pulse position and sign information. The MVDR spectrum below is used to determine the gains of the pulses

that the first and third pulses are negative (i.e. $s_1 = s_3 = -1$), and the true pulse positions are $(5, 7, 8, 12, 30)$.

The MVDR spectrum of order $2K - 1 = 9$ is constructed, and is sampled at the positions of the $2K = 10$th order Linear Prediction filter's zeros to determine gain information $(1, 10, 7, 4, 3)$. Using the sign information from the Linear Prediction filter, we determine pulse amplitudes of $(-1, 10, -7, 4, 3)$.

In summary, the LP filter's zero location property is used to locate the pulse positions and obtain the sign information, and the MVDR spectrum's accurate amplitude modeling is used to obtain the pulse gains. In terms of practical systems, the reflection coefficients need to be accurately quantized in order to preserve the pulse location information and relative amplitude information.

4. REFERENCES

[1] W.B. Kleijn and K.K. Paliwal, Editors. *Speech Coding and Synthesis*. Elsevier, 1995.

[2] E. Denoel and J-P. Solvay. "Linear Prediction of Speech with a Least Absolute Error Criterion". *IEEE Trans. Acoustics, Speech, and Signal Processing*, 33(6):1397-1403, Dec. 1985.

[3] J. Lansford and R. Yarlagadda. "Adaptive L_p Approach to Speech Coding" In *Proc. ICASSP '88*, New York, NY, vol. 1:335-338, April 1988.

[4] C-H. Lee. "On Robust Linear Prediction of Speech" *IEEE Trans. Acoustics, Speech, and Signal Processing*, 36(5):642-650, May 1988.

[5] Y. Li. "A Globally Convergent Method for l_p Problems" *SIAM J. Optimization*, 3(3):609-629, Aug. 1993.

[6] S.F. Cotter, M.N. Murthi, and B.D. Rao "Fast Basis Selection Methods" In *Proc. 31st Asilomar Conf. on Signals, Systems, and Computers*, Nov. 1997.

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

[8] M.N. Murthi and B.D. Rao. "Minimum Variance Distortionless Response (MVDR) Modeling of Voiced Speech". In *Proc. ICASSP 97*, Munich, Germany, April 1997.