N88-25753

# The DESIGN and PERFORMANCE of a REAL-TIME SELF EXCITED VOCODER[1]

Richard C. Rose, T. P. Barnwell III, S. McGrath

Georgia Institute of Technology
School of Electrical Engineering
Digital Signal Processing Laboratory
Atlanta, Georgia 30332
United States

## Abstract

This paper is concerned with a generic class of predictive speech coders that includes the newly proposed The Self Excited Vocoder (SEV) [5] and the well known Code Excited Linear Predictive Coder (CELPC) [6]. All members of this class form an excitation sequence for a linear predictive model filter using the same general model for the excitation signal. The general excitation model is based on a block coding technique where each sequence is drawn from an ensemble of sequences. This paper reports on two developments related to this general model. The first development is a new type of excitation ensemble that can in general be populated by many different types of sequences. The second development is a means of populating this new type of ensemble based on a vector quantizer design procedure using a new distortion measure.

## 1 Introduction

A general model for the excitation signal in linear predictive speech coders was originally presented in [5]. Formal subjective tests, summarized in [4], characterized the performance of selected coders in this general class of predictive speech coders. A Self Excited Vocoder has been implemented in real time on a single circuit board using the AT&T DSP32 floating point digital signal processing devices [1]. This implementation will serve as a prototype vocoder in the NASA sponsored Mobile Satellite Communications Project.

This paper presents a new approach to the excitation modeling problem in self excited and code excited vocoders. The paper begins by reviewing the general model for the excitation signal in this class of predictive speech coders, and introduces a new type of excitation ensemble. Then a new procedure for populating the excitation ensemble using a procedure based on an iterative vector quantizer design algorithm is discussed. Finally, the last section, a new distance measure for the vector quantization procedure is introduced.

## 2 A New Class of Excitation Ensembles

The general model for the excitation signal in this class of coders is described by the block diagram in Figure 1a. The excitation signal, $e[n]$, is a linear combination of component excitation sequences, $e_k[n]$, where the $k$th sequence is chosen from the associated excitation ensemble, $\mathcal{F}_k$. An excitation ensemble is simply a collection of discrete functions, $f_\gamma[n]$, indexed in sample space by $\gamma$ and indexed in time by $n$. The optimum ensemble index,

[1] This work was supported by the Jet Propulsion Laboratory under contract number 957074

$\gamma_k$, and gain, $\beta_k$, associated with the $k$th excitation sequence are found by exhaustively searching through the excitation ensemble, $\mathcal{F}_k$, for that ensemble function that minimizes a weighted mean squared error [5].

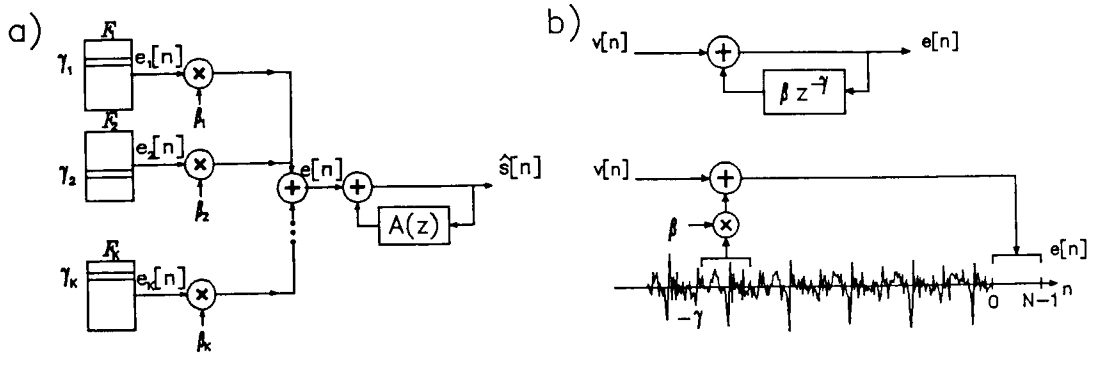

**Figure 1:** *a)* **Model of the excitation signal for a generic class of predictive speech coders.** *b)* **Ensemble search interpretation of a single tap long–term predictor.**

Examples of some existing predictive coders can be identified if the excitation ensemble is constrained to contain a particular class of sequences. For example, the well known CELPC chooses an optimum excitation sequence from a stochastic ensemble, where each ensemble sequence is populated by Gaussian random varieties [6]. Figure 1b shows how a simple long–term predictor can be interpreted as a time–varying excitation ensemble. In this case, the ensemble is the memory of a long–term predictor, whose predictor delay can vary over the expected range of a pitch period in speech. Each ensemble sequence is formed by sliding an N point rectangular window along the memory of the long–term predictor. The optimum ensemble sequence corresponds to an $N$ point sequence beginning at sample $-\gamma$ in the memory of the long–term predictor. This type of ensemble, referred to here as the "self excitation" ensemble, forms the basis for the SEV. After a brief period of initialization, the SEV derives its excitation signal, $e[n] = \beta e[n-\gamma]$, solely from this type of ensemble.

The flexibility of the most general model of the excitation signal is derived from the fact that it poses no structure on the functions contained in the excitation ensemble. From the model definition, there is no fundamental requirement that an excitation ensemble be homogeneous. Thus, a single excitation ensemble can contain more than one class of sequences. For example, an ensemble can be formed by combining a set of time–varying sequences chosen from the memory of a long–term predictor with a set of fixed Gaussian random sequences. Figure 2a illustrates an interpretation of a simple coder whose excitation is derived from this type of ensemble. While the figure suggests that a hard classification procedure is taking place, this is actually not the case. The ensemble search procedure chooses a single sequence from the entire ensemble, so the determination of which class of sequences is used is made by choosing the single sequence which results in the least measured distortion. This type of excitation ensemble will be referred to as a nonhomogeneous ensemble, and can, in general, contain many different classes of sequences. The particular ensemble illustrated by the block diagram in Figure 2a is described by the excitation signal, $e[n] = \beta z_\gamma[n]$, where

$$z_\gamma[n] = \begin{cases} v_\gamma[n] & 1 \le \gamma \le C \\ e[n-\gamma] & C < \gamma \le F \end{cases}, \tag{1}$$

and the fixed sequences, $v_\gamma[n]$, may be populated in many different ways.

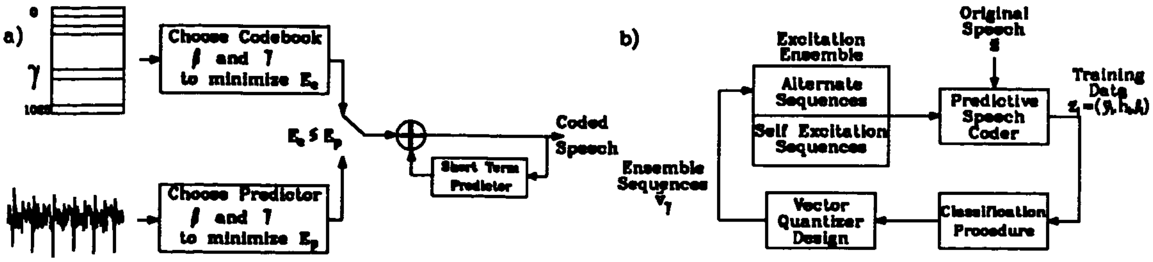

**Figure 2:** *a)* **An interpretation of a simple nonhomogeneous predictive speech coder.** *b)* **Block diagram illustrating a procedure for determining the fixed ensemble sequences in a nonhomogeneous excitation ensemble.**

## 3 Populating Nonhomogeneous Ensembles

This section describes a technique for determining the fixed sequences, $v_\gamma[n]$, in Equation 1 using the vector quantizer design procedure of Linde et al [2]. Following the reasoning of Davidson et al, the distance measure used for the vector quantization procedure can be the same weighted mean squared distance used for coding the excitation signal in this class of coders [7]. The following discussion describes the vector quantizer design procedure as it applies to populating the fixed sequences of the nonhomogeneous excitation ensemble.

The generalized Lloyd algorithm, originally introduced in [2], is an iterative algorithm for designing an optimum vector quantizer by a method of successive approximation. The vector quantizer design procedure determines the the sequences, $v_\gamma[n]$, $\gamma = 1, \ldots, C$, of Equation 1 from the training vectors, $\vec{z}_i$, $i = 1, \ldots, n$, derived from the original speech. At each iteration of the algorithm the training vectors are partitioned into clusters, and cluster centroids are computed based on the partitioning of the data. The splitting algorithm of Linde et al is used here to provide the initial cluster centroids. The cluster centroids that exist upon termination of the algorithm form the resulting excitation ensemble.

Figure 3 is a block diagram illustrating the computation of the distortion, $d(\vec{v}_\gamma, \vec{z}_i)$, that is used for the vector quantizer data set partitioning and clustering procedures. For each excitation analysis frame, $i$, the coder represents the residual vector, $\vec{r}_i$, with an ensemble vector, $\vec{v}_\gamma$. The coder also computes the short-term predictor, $A_i(z)$, and the excitation gain, $\beta_i$. The Atal LPC based weighting filter, $W_i(z)$ [6], is used to compute the weighted Euclidean distance. The distance between training vector, $\vec{z}_i$, and ensemble vector, $\vec{v}_\gamma$ can be expressed as

$$d(\vec{z}_i, \vec{v}_\gamma) = \sum_{n=0}^{N+L-2} \left( y_i[n] - \beta_i \sum_{l=0}^{N-1} v_\gamma[n] h_i[n-l] \right)^2 , \tag{2}$$

where $h_i[n]$ is a finite length impulse response approximation to the cascaded synthesis and error weighting filters in Figure 3. The length of this impulse response is approximated as $L$ samples ($L \approx 10$). The distance calculation in Equation 2 suggests the form of the training data required for each excitation frame. To compute this distance for the $i$th excitation frame, the weighted speech $\vec{y}_i$, the impulse response $\vec{h}_i$, and the ensemble gain $\beta_i$ must all be derived from the input speech. The form of each training vector is then given as $\vec{z}_i = \left(\vec{y}_i, \vec{h}_i, \beta_i\right)$ . Therefore, the training data is derived from the original speech using the

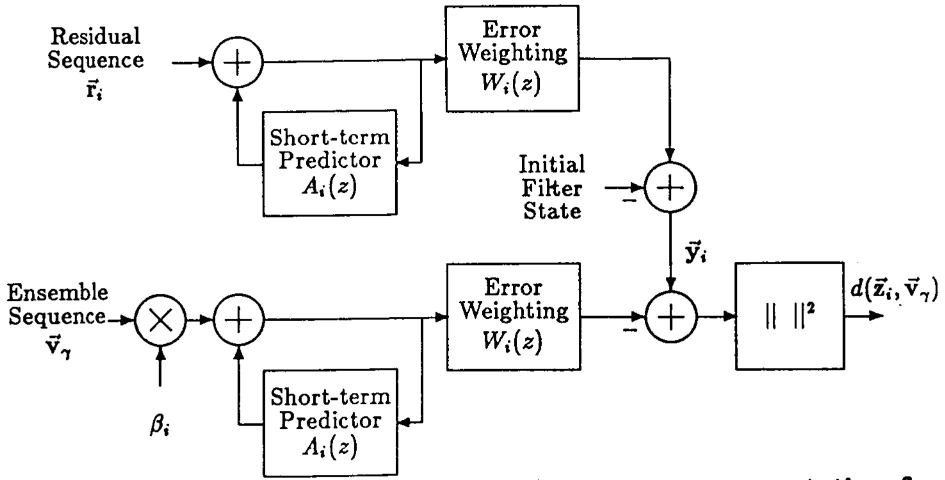

**Figure 3: Block diagram illustrating the distortion measure computation for Equation 2.**

predictive speech coder itself. The specification of an initial excitation ensemble for this coder is necessary for the generation of the training data.

In this research, a nonhomogeneous ensemble was generally divided into *self excitation sequences* and *alternate sequences* The procedure for populating the alternate sequences of the nonhomogeneous ensemble shown in Figure 2a is illustrated by the block diagram shown in Figure 2b. The procedure begins by generating the training data, $\vec{z}_i$, using an predetermined set of signals for the alternate sequences. In this research, the alternate sequences are populated by independent Gaussian random varieties. Once the training data has been generated, a classification procedure is used to select a subset of the training data to be used as input to the vector quantizer design procedure. This classification procedure simply chooses those training vectors where the predictive speech coder provides a poor representation of the original speech. Finally, the vector quantizer design procedure produces sequences that are used to populate the alternate sequences in the nonhomogeneous ensemble. This procedure is described in the next section.

## 4 A New Vector Quantizer Distortion Measure

This section describes a new distance measure for use in the iterative excitation vector quantization procedure. The new distance measure follows immediately from Equation 2, and results in circularly defined excitation ensemble sequences. The discussion is broken into three parts. First, the centroid calculation following from the weighted Euclidean distance measure of Equation 2 is described. Second, the short–comings of this distance measure when applied to vector quantization of the excitation signal are discussed. Finally, the new distance measure is introduced.

Determining the centroid for a given cluster of training vectors corresponds to finding that sequence, $\vec{v}_\gamma$, that minimizes an average distortion for the distance measure in Equation 2. This average distortion represents an average over all of the training vectors belonging to the cluster. Minimizing the average error for a cluster containing $M$ training vectors with respect to $v_\gamma[k]$, $k = 0, \ldots, N-1$, yields the matrix equation

$$\sum_{i=1}^{M} \vec{\mathbf{q}}_i = \sum_{i=1}^{M} \mathbf{R}_i \vec{\mathbf{v}}_\gamma \, . \tag{3}$$

The vector $\vec{\mathbf{q}}_i$ is an $N$ length vector where the $k$th element corresponds to the crosscorrelation between the weighted speech and the impulse response for excitation frame $i$,

$$q_i[k] = \beta_i \sum_{n=0}^{N+L-2} y_i[n] h_i[n-k], \; k = 0, \ldots, N-1 \; . \tag{4}$$

The matrix $\mathbf{R}_i$ is an $N \times N$ toeplitz matrix where the element in the $l$th row and $k$th column is given by the impulse response for excitation frame $i$,

$$R_i[l,k] = \beta_i^2 \sum_{l=0}^{N-1} h_i[n-l] h_i[n-k] \; . \tag{5}$$

The matrix order, $N$, corresponds to the length of the excitation analysis frame, which is typically about twenty samples. Hence, computing the cluster centroid is not a computationally expensive procedure, requiring only the solution of a twentieth order Toeplitz matrix equation.

A major shortcoming of the above algorithm concerns the weighted Euclidean distance given in Equation 2. By this measure, the distance between two training vectors, where both vectors represent very similar excitation signals, may actually be very large. This is due to the fact that the excitation analysis window is placed asynchronously with respect to any significant events that may occur in the excitation signal. About 24,000 training vectors derived from isolated words uttered by a single speaker were used as training data for this algorithm. Ensemble sequences derived from this algorithm were used to code a short utterance from the same speaker. There was a significant improvement in segmental signal–to–noise ratio using this new ensemble over that of a Gaussian ensemble. However, the improvement in subjective performance was not significant when judged by the authors in informal listening tests. A modification to this procedure is proposed here that reduces the dependency of the training vectors on the position of the associated excitation analysis frame. The modification to the design procedure results in a redefinition of the distance measure and centroid calculation of the vector training algorithm.

The modification is based on simple permutations of the weighted speech that is used to form the training vector $\vec{\mathbf{z}}_i$. The vector valued permutation $\pi_k$ is a $k$ sample circular right shift,

$$\pi_k(\vec{\mathbf{y}}) = (y[N-k], y[N-k+1], \ldots, y[0], y[1], \ldots, y[N-k-1]) \; . \tag{6}$$

By applying one of the permutations, $\{\vec{\pi}_k : k = 0, \ldots, N-1\}$, to the training data, similar events occurring in different excitation frames may be aligned in time.

The distance measure and centroid calculation can be modified to exploit this behavior. First, the $k$th permutation of the $i$th training vector is defined as $\vec{\pi}_k^i(\vec{\mathbf{z}}_i) = \left(\vec{\pi}_k^i(\vec{\mathbf{y}}_i), \vec{\mathbf{h}}_i, \beta_i\right)$. The weighted Euclidean distance of Equation 2 is restated as

$$d(\vec{\mathbf{z}}_i, \vec{\mathbf{v}}_\gamma) = \min_{\vec{\pi}_k^i \, : \, k=0,\ldots,N-1} d(\vec{\pi}_k^i(\vec{\mathbf{z}}_i), \vec{\mathbf{v}}_\gamma) \; . \tag{7}$$

The distance between a training vector and a cluster centroid is therefore defined as the minimum weighted Euclidean distance across all possible permutations of the input data. Having found the optimum partition by minimizing the average distortion, the centroid vector, $v_\gamma$, for centroid, $\gamma$, can be determined by solving the matrix equation,

$$\sum_{i=1}^{M} \vec{\pi}_k^i(\vec{\mathbf{q}}_i) = \sum_{i=1}^{M} \mathbf{R}_i \vec{\mathbf{v}}_\gamma \; . \tag{8}$$

In this equation, $\vec{\pi}_k^i$ is the optimum permutation for training vector $i$, and $M$ is the total number of training vectors in the cluster.

## 5 Conclusions

This paper has introduced the nonhomogeneous excitation ensemble as a new type of ensemble used in a generic class of predictive speech coders. An iterative vector quantization procedure using a newly defined distance measure has been discussed as a means for populating the sequences for a specific nonhomogeneous ensemble. In this new procedure, the optimum choice of ensemble sequence is less dependent on the alignment of the excitation analysis frame with the original speech waveform. The procedure involves applying a set of circular permutations to the training data in order to time align similar events in different training vectors. The ensemble search procedure for this newly defined ensemble involves an exhaustive search, computing the weighted mean squared coding error for each circular permutation of each $N$ point ensemble sequence. This is essentially equivalent to increasing the number of sequences in the ensemble from $F$ sequences to $FN$ sequences. However, the number of operations required to search this ensemble can be considerably reduced by using the recursive ensemble search procedure introduced in [3].

## References

[1] T. P. Barnwell III, R. C. Rose, and S. McGrath. A real-time implementation of a 4800 bps. self excited vocoder using the AT & T WE-DSP32 signal processing microprocessor. *Proc. Speech Technology Conf.*, Apr. 1987.

[2] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizor design. *IEEE Trans. Commun.*, COM-28:84–95, Jan. 1980.

[3] R. C. Rose and T. P. Barnwell III. The design and performance of an effective class of predictive speech coders. *Submitted for Review IEEE Trans. Acoust., Speech, Signal Processing*, June 1987.

[4] R. C. Rose and T. P. Barnwell III. Quality comparison of low complexity 4800 bps. self excited and code excited vocoders. *Proc. Inter. Conf. on Acoustics, Speech, and Signal Proc.*, April 1987.

[5] R. C. Rose and T. P. Barnwell III. The self excited vocoder–an alternate approach to toll quality at 4800 bps. *Proc. Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 453–456, April 1986.

[6] M. R. Schroeder and B. S. Atal. Code excited linear prediction: high quality speech at very low bit rates. *Proc. Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 937–940, April 1985.

[7] G. Davidson M. Yong and A. Gersho. Real-time vector excitation coding of speech at 4800 bps. *Proc. Inter. Conf. Acoust., Speech, Sig. Processing*, 51.4.1–51.4.4, 1987.