EFFICIENT CODING OF LPC PARAMETERS BY TEMPORAL DECOMPOSITION

Bishnu S. Atal

Bell Laboratories

Murray Hill, New Jersey 07974

ABSTRACT

This paper describes a method for efficient coding of LPC log area parameters. It is now well recognized that sample-by-sample quantization of LPC parameters is not very efficient in minimizing the bit rate needed to code these parameters. Recent methods for reducing the bit rate have used vector and segment quantization methods. Much of the past work in this area has focussed on efficient coding of LPC parameters in the context of vocoders which put a ceiling on achievable speech quality. The results from these studies cannot be directly applied to synthesis of high quality speech. This paper describes a different approach to efficient coding of log area parameters. Our aim is to determine the extent to which the bit rate of LPC parameters can be reduced without sacrificing speech quality. Speech events occur generally at non-uniformly spaced time intervals. Moreover, some speech events are slow while others are fast. Uniform sampling of speech parameters is thus not efficient. We describe a non-uniform sampling and interpolation procedure for efficient coding of log area parameters. A temporal decomposition technique is used to represent the continuous variation of these parameters as a linearly-weighted sum of a number of discrete elementary components. The location and length of each component is automatically adapted to speech events. We find that each elementary component can be coded as a very low information rate signal.

INTRODUCTION

A long standing goal of speech research has been to develop a simple and efficient description of speech events. Such a description is important for many practical applications, such as speech coding, speech synthesis, and speech recognition. For example, in speech coding, our aim is to represent the speech wave by a small number of time-varying parameters which are capable of regenerating speech at low bit rates without significant distortion. Speech wave has a bandwidth of about 4 kHz. Speech parameters, such as a log area parameter determined by LPC analysis [1-4], can be limited in bandwidth to about 50 Hz without introducing any additional distortion due to band limiting [3,4]. The total bandwidth for 12 log area parameters is therefore 600 Hz, which is considerably lower than 4000 Hz required for the speech signal. A major source of redundancy in LPC area parameters arises from the correlations between successive time frames. These correlations are caused by a number of factors involved in human speech production. Most obvious of these is the smooth movement of different articulators in the vocal tract.

A common method of coding log area parameters is time sampling followed by scalar or vector quantization [5,6]. If each parameter is band limited to 50 Hz, it can be sampled at 100 Hz without loss of information. Scalar quantization of each frame of log area parameters typically requires 48 bits which yields a bit rate of 4800 bits/sec. What can be done to reduce this bit rate? One possibility is to reduce the bandwidth of each parameter even more. For example, if the bandwidth is lowered to 25 Hz, the parameters can be sampled at 50 Hz yielding a bit rate of 2400 bits/sec. However, a bandwidth of 25 Hz is usually too small to represent fast variations of short transient sounds accurately.

Speech events occur generally at non-uniformly spaced time intervals. Moreover, articulatory movements for some speech sounds are fairly slow while for others they are relatively fast. Uniform sampling of speech parameters is thus not efficient. With uniform sampling, one is forced to use a small sampling interval to be able to represent the fastest speech event accurately. Non-uniform sampling of speech parameter variations is in general more efficient because the sampling interval can be adapted to the nature of speech events. Since speech sounds are produced in human speech at an average rate of approximately between 10 and 15 sounds/sec, it should be sufficient to specify the acoustic parameters at an average rate of less than 15 frames/sec. In this paper, we present a procedure to break up the continuous variation of log area parameters into discrete units of variable lengths located at non-uniformly spaced time intervals. Coding efficiency is achieved by coding these units rather than the parameters themselves.

TEMPORAL DECOMPOSITION MODEL FOR LOG AREA PARAMETERS

Consider the variation of log area parameters as a function of time. Let $y_i(n)$ be the ith log area parameter at the nth sampling instant. It is assumed that the parameters have been sampled at closely spaced time intervals small enough to represent accurately even the fastest speech events. The sampling interval is typically 1 to 2 msec. The index i varies from 1 to p where p is the total number of area parameters determined by LPC analysis. The value of p is typically 16 for speech sampled at 8 kHz. The index n varies from 1 to N where $n=1$ is the first sample in the utterance and $n=N$ is the last sample in the utterance. Figure 1 shows the first 8 log area parameters for the utterance "Joe brought a young girl" spoken by a male speaker. The rms amplitude is also shown on the figure.

We represent $y_i(n)$ as

$$\hat{y}_i(n) = \sum_{k=1}^{m} a_{ik}\phi_k(n), \quad 1 \leqslant n \leqslant N, \quad 1 \leqslant i \leqslant p, \tag{1}$$

where $\hat{y}_i(n)$ is the approximation of $y_i(n)$ produced by the model, $\phi_k(n)$ is the kth interpolation function at the sampling instant n, and a_{ik} is the contribution of the kth interpolation function to the ith area parameter. The value of m corresponds roughly to the number of speech (and silence) events in the speech utterance in the time interval $n=1$ to $n=N$.

Equation (1) can be expressed in matrix notations as

$$Y = A\ \Phi \tag{2}$$

where Y is a $p \times N$ matrix whose (i,n) element (ith row and nth column) is $y_i(n)$, A is a $p \times m$ matrix whose (i,k) element is a_{ik}, and Φ is a $m \times N$ matrix whose (k,n) element is $\phi_k(n)$. We wish to determine matrices A and Φ so that the bit rate required to represent them is minimum.

2.6

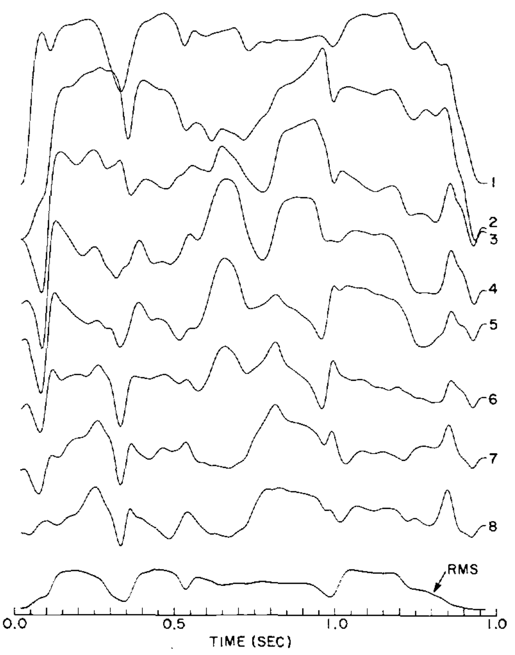

Fig. 1. Plot of first 8 log area parameters and rms amplitude as a function of time for a sentence-length utterance, "Joe brought a young girl", spoken by a male speaker.

We will assume that the functions $\phi_k(n)$ are ordered with respect to their locations in time. That is, the function $\phi_2(n)$ occurs later than the function $\phi_1(n)$ and so on. Each $\phi_k(n)$ is supposed to correspond to a particular speech event. Since a speech event lasts for a short time, each $\phi_k(n)$ should be non zero only over a small range of values of n. A typical $\phi(n)$ is sketched in Fig. 2. For efficient coding, the matrix Φ should be a sparse matrix.

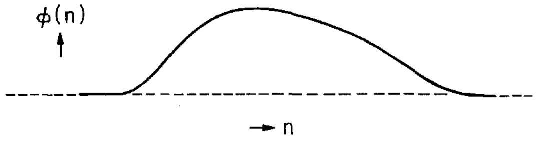

Fig. 2. Idealized sketch of a typical interpolation function.

We illustrate the above point in the example shown in Fig. 3. We show there three functions of time $y_1(n)$, $y_2(n)$, and $y_3(n)$. These functions were constructed by combining the three functions of time $\phi_1(n)$, $\phi_2(n)$, and $\phi_3(n)$, shown in Fig. 3(b), using three different sets of coefficients a_{ik} in Eq. (1). Thus, all of the $y(n)$ of Fig. 3 follow Eq. (1) exactly. Since each $\phi(n)$ is limited to a much shorter interval in comparison to any one of the $y(n)$ and the bandwidth of each $y(n)$ is the maximum bandwidth of any one of the $\phi(n)$, it is obvious that direct coding of $y(n)$ will take more bits than the coding of the $\phi(n)$ and the coefficients used to combine $\phi(n)$ to form $y(n)$.

As mentioned earlier, the value of m in Eq. (1) is related to the duration of the speech segment and the number of sounds the speech segment contains. In general, m is proportional to N. Consider a short segment of speech such that the rank of the matrix $Y \geqslant m$. The maximum rank of the matrix Y is p, no matter how long the speech segment. Previous work suggests that the rank of Y is about 10 even for very long utterances. To satisfy the requirement that rank of the matrix $Y \geqslant m$, the duration of speech segment should be

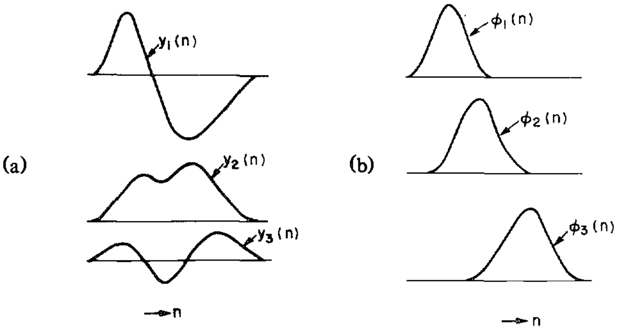

Fig. 3. (a) Three different linear combinations of the basis functions shown on the right and (b) the basis functions.

approximately 0.2 to 0.3 sec. Whenever the rank of $Y \geqslant m$, Eq. (2) can be inverted to yield

$$\Phi = (A^t A)^{-1} A^t Y, \tag{3}$$

which implies that

$$\phi_k(n) = \sum_{i=1}^{p} w_{ki} y_i(n), \quad 1 \leqslant k \leqslant m, \; 1 \leqslant n \leqslant N, \tag{4}$$

for some choice of the weights w_{ki}. That is, each interpolation function ϕ is a linear combination of the y's.

The problems related to determining the rank of Y are easily resolved by looking at the eigenvalues obtained from the singular-value decomposition of Y. We represent Y as

$$Y^t = U \, D \, V^t, \tag{5}$$

where U is a $N \times p$ orthogonal matrix, V is a $p \times p$ orthogonal matrix, D is a diagonal matrix of eigenvalues, and the superscript t on a matrix means its transpose. Typical values of the first ten eigenvalues, for a short speech segment 0.25 sec in duration, are 0.83, 0.52, 0.16, 0.13, 0.06, 0.03, 0.03, 0.02, 0.01, and 0.01, respectively. Assuming that an error of 0.05 in log areas is insignificant, the rank of Y is 5. We then set m to 5.

It is obvious from Eqs. (4) and (5) that an interpolation function $\phi_k(n)$ can also be represented as

$$\phi_k(n) = \sum_{i=1}^{m} b_{ki} u_i(n), \tag{6}$$

where $u_i(n)$ is the element in the nth row and the ith column of the matrix U and b_{ki} are a set of amplitude coefficients.

DETERMINATION OF INTERPOLATING FUNCTIONS

We define a measure of distance of $\phi(n)$ from the sample $n = l$ as

$$\theta(l) = [\sum_n (n-l)^2 \phi^2(n) \,/\, \sum_n \phi^2 n]^{1/2}, \tag{7}$$

where the sum over the index n extends over the duration of the speech segment. The optimum $\phi(n)$ is chosen so as to minimize the distance function $\theta(l)$.

Minimization of $\theta(l)$

Since the problem of minimizing $\theta(l)$ is equivalent to the problem of minimizing $\ln \theta(l)$, we set the derivatives of $\ln \theta(l)$ with respect to the unknown amplitude coefficients b_{ki} of Eq. (6) equal to zero. We then obtain

$$\sum_n (n-l)^2 \frac{\partial}{\partial b_r} \phi^2(n) = \lambda \sum_n \frac{\partial}{\partial b_r} \phi^2(n), \quad 1 \leqslant r \leqslant m, \tag{8}$$

where

$$\lambda = [\sum_n (n-l)^2\phi^2(n) \,/\, \sum_n \phi^2(n)] = \theta^2{}_{\min}. \tag{9}$$

From Eq. (6), we can write

$$\phi^2(n) = \sum_{i=1}^{m}\sum_{j=1}^{m} b_i b_j u_i(n) u_j(n), \tag{10}$$

where the subscript k has been dropped. Then,

$$\frac{\partial}{\partial b_r}\phi^2(n) = 2\sum_{i=1}^{m} b_i u_i(n) u_r(n), \quad 1 \leqslant r \leqslant m. \tag{11}$$

On combining Eqs. (8) and (11), one obtains

$$\sum_{i=1}^{m} b_i \sum_n (n-l)^2 u_i(n) u_r(n) = \lambda \sum_{i=1}^{m} b_i \sum_n u_i(n) u_r(n) = \lambda b_r. \tag{12}$$

Equation (12) can be expressed in matrix notations as

$$R\mathbf{b} = \lambda\mathbf{b}, \tag{13}$$

where the element in the ith row and rth column of the matrix R is given by

$$R_{ir} = \sum_n (n-l)^2 u_i(n) u_r(n). \tag{14}$$

Equation (13) has exactly m solutions. If all the λ's are different, the solution corresponding to the smallest λ provides the correct **b**. In case they are not, the minimum value of λ determines the optimum **b**; although the choice of optimum **b** is not unique. The nearest $\phi(n)$ is determined from the coefficients b_i's by using Eq. (6). The location of the nearest $\phi(n)$ is given by

$$\nu(l) = [\sum_n (n-l)\phi^2(n) \,/\, \sum_n \phi^2(n)]. \tag{15}$$

The function $\nu(l)$ crosses the $\nu(l)=0$ axis from the positive side at each sampling instant l which equals the location of one of the $\phi_k(n)$ for some k.

Better estimates of $\phi(n)$'s are obtained by repeating the minimization for all values of l for which $\nu(l)=0$, and using a time interval which contains exactly 5 speech events ($m = 5$). This indeed is always possible except at the beginning or at the end of an utterance which begins or ends with a silence. A lower value of m is used in these shorter segments. The first and last $\phi(n)$'s correspond to "silence" segments.

Determination of amplitude coefficients a_{ik}

The amplitude coefficients a_{ik} of Eq. (1) are determined by minimizing the mean-squared error E defined by

$$E = \sum_n [y_i(n) - \sum_{k=1}^{M} a_{ik}\,\phi_k(n)]^2, \tag{16}$$

where M represents the total number of speech events within the range of index n over which the sum is carried out. On setting the partial derivatives of E with respect to the coefficients a_{ik} equal to zero, we obtain a set of simultaneous linear equations

$$\sum_{k=1}^{M} a_{ik} \sum_n \phi_k(n)\phi_r(n) = \sum_n y_i(n)\phi_r(n), \quad 1 \leqslant r \leqslant M, \quad 1 \leqslant i \leqslant p, \tag{17}$$

which can be solved for the unknown coefficients a_{ik}.

Iterative Refinement of $\phi_k(n)$'s and a_{ik}'s

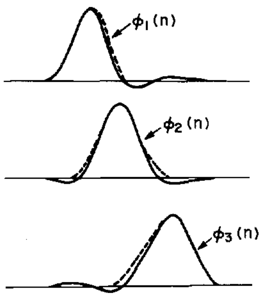

Fig. 4. Plots of the interpolation functions obtained by temporal decomposition of the curves shown in Fig. 3(a). The dashed curves are the actual basis functions illustrated in Fig. 3(b).

Figure 4 shows a plot (solid line) of the interpolation functions $\phi_k(n)$, obtained from the above procedure, for the example illustrated in Fig. 3. The actual functions $\phi_k(n)$ are also shown as dashed curve on the same plot. The agreement between the two is close except for the presence of a number of small ripples and the narrowing of the major lobe. The mean-squared criterion used for the distance function shown in Eq. (7) is a contributing factor for these differences. We discuss here an iterative refinement procedure for obtaining better estimates of $\phi_k(n)$ and a_{ik}. For a given set of a_{ik}, we determine $\phi_k(n)$ to minimize the error E given in Eq. (16). This is done by setting the partial derivatives of E with respect to $\phi_k(n)$ equal to zero. One then obtains

$$\frac{\partial E}{\partial \phi_r(n)} = 2\sum_{i=1}^{p}[y_i(n) - \sum_{k=1}^{M} a_{ik}\phi_k(n)]a_{ir} = 0, \quad 1 \leqslant r \leqslant M, \tag{18}$$

which further simplifies to

$$\phi_r(n) = [\sum_{i=1}^{p} y_i(n)a_{ir} - \sum_{k \neq r}\phi_k(n)\sum_{i=1}^{p} a_{ik}a_{ir}] \,/\, [\sum_{i=1}^{p} a^2{}_{ir}]. \tag{19}$$

Since the the coding of minor lobes of $\phi(n)$ can use a significant number of bits, we retain only the major lobe of the interpolation functions and set the functions equal to zero every where else. The resultant $\phi_k(n)$ are used again in Eq. (17) to obtain an even better estimate of a_{ik}. The procedure is repeated until the decrease in error E, as defined in Eq. (16), falls below a predetermined threshold value. Four iterations are usually sufficient to converge both a_{ik} and $\phi_k(n)$ to stable set of values.

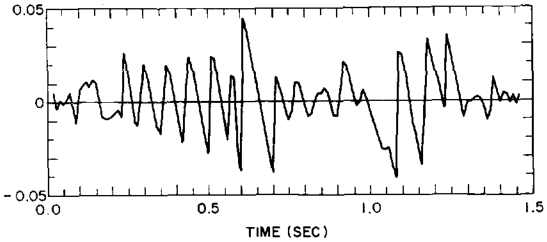

Fig. 5. Plot of the timing function $\nu(l)$ for the utterance "Joe brought a young girl" shown in Fig. 1.

RESULTS

The above procedure was carried out on several sentences spoken both by male and female speakers. We present results here for one sentence "Joe brought a young girl" spoken by a male speaker. The timing function $\nu(l)$ defined in Eq. (15) is illustrated in Fig. 5. A new value of $\nu(l)$ was computed once every 10 msec. Each zero crossing from positive to negative values indicates the location of a speech event. The zero crossings going from negative side to positive side signify a rapid shift from one $\phi(n)$ to the next. This shift is very sharp as expected. The function $\nu(l)$ has a total of 23 negative-going zero crossings. The interpolation functions $\phi(n)$ located at these time instants are shown in Fig. 6 together with the corresponding speech waveforms. As expected, the interpolation functions for short transient sounds last over a short time interval while the interpolation functions for relatively stationary vowel sounds last over a much longer time interval.

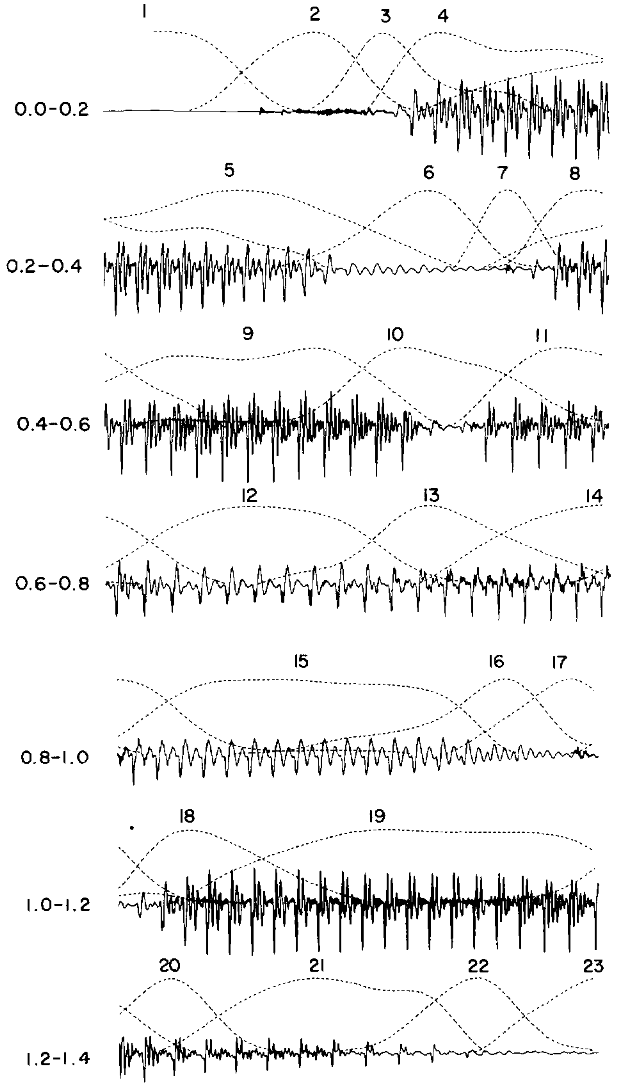

Fig. 6. Plots of speech waveform for the utterance "Joe brought a young girl" and the various interpolation functions determined by the temporal decomposition technique. The time intervals (in secs) for the different segments are marked in the left margin in each case.

Figure 7 shows the first 8 log area parameters and the rms value as a function of time for the utterance shown in Fig. 6. The solid curve shows the original areas determined by LPC analysis of the speech wave. The dashed curve shows the approximation of each $y_i(n)$ by the additive model defined in Eq. (1). The results for the remaining 8 log areas are similar. As can be seen, the agreement between the model and the actual results is very good.

Bit Rate Required to Encode Area Parameters

The interpolating functions in general vary smoothly as a function of time. We have determined the bandwidth of each interpolating function from its amplitude spectrum. An effective bandwidth for each spectrum can be defined as the frequency at which the amplitude spectrum falls to 1/20 of its value at d.c. We find that an average of 4 samples per $\phi_k(n)$ are needed to sample the function at the Nyquist rate.

It is sufficient to encode each sample of $\phi_k(n)$ at 4 bits/sample to keep the error in the log area parameters to be less than 0.10. Thus the total number of bits required to encode each $\phi_k(n)$ is 16 bits.

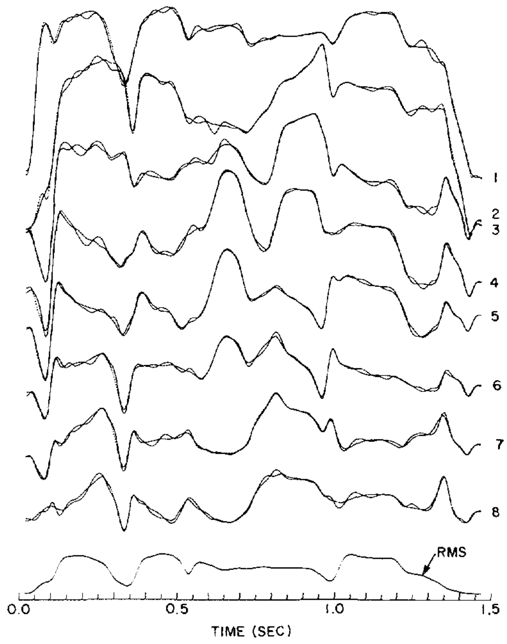

Fig. 7. Plot comparing the model-generated area parameters (dashed curve) with the actual areas (solid curve) of Fig. 1.

For each k, the coefficients a_{ik} need to be coded with the same accuracy as a single frame of log area parameters. With scalar quantization, we find that 48 bits/frame are sufficient [4]. Recent work on vector quantization suggests that number of bits/frame can be reduced even further [6].

The total information rate for encoding of log area parameters depends upon the number of speech events (or sounds) spoken per second. For slow speaking rate, this number is about 10. Assuming 5 bits to represent the location of each ϕ, the bit rate for coding both a_{ik} and $\phi_k(n)$ will then be $(48 + 16 + 5) = 690$ bits/sec. The bit rate would increase to 1035 bits/sec for a speaking rate of 15 sounds per sec.

REFERENCES

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

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs: Prentice Hall, 1973.

[3] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction," *J. Acoust. Soc. Amer.* vol.50, pp. 637-655, Aug. 1971.

[4] B. S. Atal, "Predictive Coding of Speech at Low Bit Rates," *IEEE Trans. Commun.* , vol. COM-30, pp. 500-614, April 1982.

[5] R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," *IEEE Trans. Acoust. Speech, Signal Processing,* vol. ASSP-23, pp. 309-321, June 1975.

[6] D. Y. Wong and B. H. Juang, "Voice Coding at 800 BPS and Lower Data Rates with LPC Vector Quantization," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing.* Paris, France, 1982, pp. 606-609.