Low Rate Audio Compression using Parametric Spectral Modeling Techniques

Udaya Bhaskar

COMSAT Laboratories
22300 COMSAT Drive
Clarksburg, MD 20871,USA

ABSTRACT

As compression rates for audio signals approach the 1 bit/sample/Hz threshold, the applicability of parametric spectral modeling techniques is of increasing interest. This paper reports continued work in this area, in developing further the technique of adaptive predictive coding with transform domain quantization (APC-TQ). In particular, improvements in the efficiency of representation of the signal spectrum representation and the modeling of harmonic structure are addressed.

1.0 Introduction

The technique of Adaptive predictive coding with transform domain quantization (APC-TQ) has been previouly reported for low rate audio compression [1, 2]. In APC-TQ, parametric spectral modeling is combined with transform domain quantization of the prediction residual waveform to encode audio signals of 5 kHz bandwidth at a rate of 17 kbit/s. This paper reports continued work in further developing this technique, with the objective of further reducting its bit rate and improving its audio quality. The work has focused on overcoming some of the limitations of the APC-TQ technique. Two such limitations which are addressed in this paper are (i) modeling of the harmonic spectral characteristics, and (ii) efficiency of spectral model parametrization. A summary of the APC-TQ technique is presented first.

1.1 The APC-TQ Technique

Two aspects of auditory signals and their perception by the human ear form the basis of audio compression. These are: (i) the non-uniform spectral power distribution that is usually a characteristic of audio signals, and (ii) the variation in perceptual sensitivity of the human auditory system across the frequency band, as a function of this spectral power distribution. Significant reductions in bit rate can be achieved by exploiting these characteristics of the signal and the human ear. This forms the basis of a number of audio coding techniques such as perceptual transform coders and subband coders, which may be viewed as non-model based techniques.

In the APC-TQ technique, time-domain prediction modeling and filtering is employed to realize a coding gain (in the form of a prediction gain) due to the non-uniform power spectral distribution. The resulting prediction residual signal is quantized in the sinusoidal frequency domain according to an adaptive bit allocation. The bit allocation is non-uniform across the frequency band, so as to exploit the perceptual sensitivity of the human auditory system.

In general, forward adaptive predictive methods such as APC-TQ can exploit the non-uniform power spectral distribution of the input signal more efficiently than transform or subband coders. In these methods, the predictors are optimized for each block of input signal samples resulting in a highly decorrelated prediction residual signal. In contrast, practical transform coding with a fixed suboptimal transform such as the discrete cosine transform (DCT) results in less complete decorrelation of the transform coefficients. In the case of subband coders, the number of subbands limits the extent to which the spectral variations are exploited. The strength of transform and subband coding schemes, which is the ease with which auditory characteristics are exploited, is retained in APC-TQ due to quantization in the transform domain. This permits direct implementation of auditory noise masking models [ref. 3], thereby maximizing the perceived quality of the reconstructed audio signal.

1.2 Codec Structure

The structure of the APC-TQ encoder is illustrated in Figure 1. The decoder is a simpler sub-set of the encoder. The encoder consists of 4 primary

1058-6393/95 $4.00 © 1995 IEEE

functional units: (i) block size adaptation, which adjusts the block size according to the quasi-stationarity of the audio signal, (ii) short term and long term prediction and filtering to translate spectral envelope and harmonic structures into coding gain, (iii) adaptive bit-allocation which distributes the available bits according to objective and perceptual criteria and (iv) transform domain quantization which digitizes the transform coefficients of the prediction residual signal.

1.2.1 Block size Adaptation

The APC-TQ encoder processes the audio samples in blocks whose size N is dynamically adapted based on the stationarity of the audio signal. The adaptative block size allows exploitation of extended periods of stationarity, and yet permits tracking of rapid changes. Block size is adapted based on a spectral distortion measure. In the present implementation, four block sizes are possible: 256, 512, 768 and 1024 samples.

1.2.2 Linear Predictive Spectral Modeling and Filtering

The power spectrum of each block of audio samples is modeled by linear predictive models [4]. The model is a product of two terms, a short term model, which models the coarse or envelope spectral variations, and a long term model, which models fine or harmonic spectral variations. The power spectrum model that results is used for the computation of the auditory noise masking threshold function for the block, which is used to determine the bit-allocation. In addition, the short term and long term predictive models are used to filter the signal, resulting in a highly uncorrelated, noise-like prediction residual signal. Subsequently, the residual signal is quantized in the transform domain and transmitted to the decoder.

Short Term Prediction Analysis and Filtering

Short term prediction is performed by predicting each sample by a weighted sum of the M samples immediately preceding it. The parameter M is the order of short term prediction, is selected based on the block size (N). Corresponding to each of the four possible block sizes, the order of short term prediction is 16, 32, 48 and 64 respectively. The weights, known as linear predictive coding (LPC) parameters, are optimized for each block by a stabilized covariance analysis method [5]. The LPC parameters are quantized and transmitted to the decoder and are also used to perform the short term prediction filtering operation. This has the effect of removing short term correlations from the input signal, resulting in the short term prediction error signal.

Long Term Prediction Analysis and Filtering

Long term prediction is based on a single long term prediction delay which is optimal for the prediction of each sample by a weighted sum of the three samples located around the delay. The optimum delay is the location of the peak of the autocorrelation function of the short term prediction error within the delay range 20-256. The long term prediction filter removes the long term redundancies from the short term prediction error resulting in the residual signal.

1.2.3 Transform Domain Quantization of the Residual

The residual signal is a highly decorrelated Gaussian white noise-like signal. It is necessary to quantize and transmit this signal to the decoder to correct for the modeling errors in the short and long term predictors, as well as provide phase information, since the predictor models are minimum phase. In APC-TQ, residual quantization is performed in the discrete cosine transform (DCT) domain, based on an adaptive non-uniform allocation of available bits.

Bit-Allocation and Quantization

The B available bits are non-uniformly allocated among the N transform coefficients based on the power spectral distribution of the input signal and perceptual noise masking criteria. The bit-allocation is performed in two steps. In the first step, 0.7B bits are allocated based purely on minimum mean squared error criteria, i.e., to minimize the reconstruction noise power. This step is necessary to ensure stable operation. In the second step, the remaining 0.3B bits are allocated based on perceptual criteria, i.e, to mainitain the reconstruction noise below or close to the auditory noise masking threshold [3]. Based upon the bit-allocation, the DCT coefficients are quantized by scalar Max quantizers optimized for zero mean, univariate Gaussian distribution.

2.0 Improvements in Harmonic Spectral Modeling

A major limitation of the current APC-TQ technique lies in the fact that only a single periodicity is modeled by the long term prediction process, which employs a single long term lag parameter. In cases where the audio signal contains a multiplicity of periodicities, this process is able to model only the dominant periodicity. The secondary periodicities are contained in the residual signal, which is encoded by the transform domain quantization process. However, since bits are not specifically allocated at the secondary

harmonic frequencies, this approach is not very efficient and in many cases results in reduced audio quality.

In order to fully exploit the harmonic spectral structure of the general class of audio signals in a parametric fashion, it is necessary that the compression technique be capable of identifying a multiplicity of harmonic frequency sets. Since spectral power tends to be concentrated at these harmonics, parametrization of these frequencies and the spectral peak behaviour at these frequencies can encode important perceptual components of the audio signal very efficiently.

As a direct implementation of the above idea, the following approach has been developed. The power spectral density (PSD) of the short term prediction error signal is computed using a high order discrete fourier transform (DFT). Typically, for a 256 sample audio signal block, a 1024 point DFT is used with a 256 point Hamming window. Let the discrete PSD of the short term prediction error signal be denoted by

$$\{S(k), 0 \le k \le \frac{N_{DFT}}{2}\},$$

where, N_{DFT} is the size of the DFT. The average PSD S_{avg} across the entire audio band is determined as:

$$S_{avg} = \frac{1}{N_{DFT}} \sum_{k=0}^{N_{DFT}-1} S(k).$$

The harmonic frequencies in the audio signal are identified by locating the local peaks in the PSD. A local peak is assumed to exist at the discrete frequency k_m if the following two simple rules are satisfied:

$$(i)\ S(k) \le S(k_m),\ \textit{for any } k \textit{ such that } |k_m - k| < 6,$$
$$\textit{and } 0 \le k \le \frac{N_{DFT}}{2}$$

and

$$(ii)\ S(k_m) > 2S_{avg}.$$

The set of all local peaks of the audio PSD is then subdivided into sub-sets, each of which contains a fundamental and its harmonics. Each discrete frequency within the range of likely fundamental frequencies, at which a local peak is displayed, is potentially a fundamental frequency. In order that a candidate frequency be selected as a fundamental, the following criteria must be satisfied:

(i) If k_1 is a candidate fundamental, it is ideally expected that local peaks also exist at its harmonics, i.e., at $\{nk_1, 1 < n < \frac{N_{DFT}}{2k_1}\}$. In reality, due to the limited accuracy of the k_1 itself, harmonics can not be expected to occur at exact multiples of the estimated fundamental frequency. Moreover, since even a small error in k_1 is amplified n times in nk_1, the above criterion must be revised in actual practice. Assuming that the $(n-1)^{th}$ harmonic has been detected at k_{n-1}, the n^{th} harmonic is assumed to exist if a peak frequency k_i exists that satisfies the following condition:

$$|k_{n-1} + k_1 - k_i| \le 0.1k_1.$$

(ii) If k_1 is a fundamental frequency, ideally $\frac{N_{DFT}}{2k_1}$ harmonics are expected to exist. However, in reality, not all harmonics may have been detected. If N_1 is the actual number of harmonics detected for k_1, the ratio $\frac{2N_1k_1}{N_{DFT}}$ is a measure of the likelihood of k_1 actually being a fundamental. This ratio is close to 1 for correct choices of fundamentals, and much lower for others.

(iii) The average power concentrated at the detected harmonics of the candidate fundamental k_1 may be computed as

$$S^1_{avg} = \frac{1}{N_1} \sum_{k \in K_1} S(k).$$

Here, K_1 is the set of all peak frequencies which satisfy the condition (i) above. The normalized power ratio,

$$\frac{S^1_{avg}}{S_{avg}}$$

is also a measure of k_1 being a fundamental. For a correctly chosen fundamental, this power ratio is much higher than 1. Otherwise, it is close to or less than 1. This indicator is useful in eliminating some of the problems associated with the existance of spurious local peaks.

By comparing the above two measures against predetermined threshold values, a set of fundamentals is identified to account for the local peak behaviour of the PSD. This method has been found to be better than 90% accurate in identifying correct fundamentals under certain controlled conditions. These conditions include the absence of rapid changes in spectral characteristics, fairly pronounced harmonic structure and a limited number of fundamental frequencies. Otherwise, the method frequently results in erroneous estimates. A major cause is the presence of spurious local peaks, often due to the sidelobe response of the window function employed in the computation of the PSD. Another frequent problem that has been observed is the doubling of the fundamental frequency. More sophisticated criteria are needed to overcome these problems and obtain a satisfactory overall performance.

3.0 Improvements in Quantization of Spectral Parameters

The linear predictive spectral model has been successfully applied in APC-TQ for modeling non-harmonic characteristics of the PSD of the audio signal. However, the efficiency of the quantization of the LPC parameters was limited due to numerical instabilities associated with high order LSF computation for audio signals. To overcome this limitation, a split-reflection coefficient vector quantization approach (also reported in ref. 6) has been developed. This approach is inherently more stable due to the direct computation of the reflection coefficients and the easily controlled stability afforded by their use. This approach is significantly more efficient than the approach being presently used in APC-TQ, resulting in a reduction of the number of bits required to represent the short term power spectrum by 50%. The split-reflection coefficient vector quantization approach is also only slightly sub-optimal compared to the split-LSF vector quantization approach.

4.0 Composite Modeling of the Magnitude Spectrum

In APC-TQ, the spectrum of the audio signal is modelled in three steps: (i) modeling of the envelope magnitude spectrum by the short term predictor, (ii) modeling of the harmonic magnitude spectrum by the long term predictor, and (iii) the quantized representation of the complex (i.e, magnitude & phase) spectrum of the residual signal. In a sense, the complex spectrum of the audio signal is represented as the product of these three spectral estimates. A question that naturally arises is whether it is possible to combine these three steps into a single step and thereby achieve improvements in efficiency and quality.

One possible approach to such a composite modeling of the magnitude component of the audio spectrum is to replace the three steps by single very high order LPC modeling of the audio spectrum. The order of this LPC model must be high enough to model the harmonic structure as well as the envelope structure of the audio signal PSD. The accuracy provided by this model must be better than or comparable to the accuracy provided by the current 3-step approach. Finally, a representation of the phase spectrum is necessary, since the minimum phase LPC model can not represent this aspect of the audio signal. This can be accomplished by encoding the phase component of the DFT.

Presently, in the APC-TQ method, a 256 sample block of audio signal is encoded using 425 bits. Of these, 320 bits are used for the quantization of DCT coefficients, 80 bits are used for the quantization of the 16-th order LPC model, 16 bits are used for the quantization of the long term model and 8 bits are used for power. Let us tentatively assume that about one-half of the 320 bits used for DCT quantization are necessary to obtain the same level accuracy for phase representation. In the composite approach, the remaining bits, i.e., 250 bits can be used for the high order LPC model of the magnitude spectrum. This permits an LPC order in the range 100 - 120, depending on the performance of the LPC quantization approach. There is also the possibility of using fewer bits for the quantization of phase information and more bits for the quantization of the magnitude spectrum.

In the composite approach, the LPC quantization process must take into account the perceptual sensitivity of the human ear. A segmented-reflection coefficient vector quantization approach is a likely candidate for LPC quantization. In this approach, the perceptual criteria must be employed during the selection of the optimum vector.

The above is only a sketch of the conceptual framework of the composite approach and a number of important details remain to be resolved. A major issue is whether the magnitude spectral approximation provided by the high-order LPC model can be comparable to or better than the magnitude spectral representation provided by the quantization of DCT coefficients. The adequecy of the range of orders that appear to be feasible (100 - 120), effects of under-modeling or over-modeling and computation complexity are issues that need to be studied. Nevertheless, the approach is promising enough to merit further investigation.

5. References

1. Udaya Bhaskar, "Adaptive Prediction with Transform Domain Quantization for Low Rate Audio Coding", 1991 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics.

2. Udaya Bhaskar, "Adaptive Prediction with Transform Domain Quantization using Block Size Adaptation and High Resolution Spectral Modeling", 1993 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics.

3. J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Criteria", IEEE Journal on Selected Areas in Communications, Volume 6, pp 314-323, February 1988.

4. L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Inc., Englewood Cliffs, N.J., 1978.

5. B. S. Atal, "Predictive Coding of Speech at Low Rates", IEEE Transactions in Communications, Vol. COM-30, No. 4, April 1982.

6. K. Law and C. Chan, "Split-Dimension Vector Quantization of Parcor Coefficients for Low Bit Rate Speech Coding", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 3, July 1994.

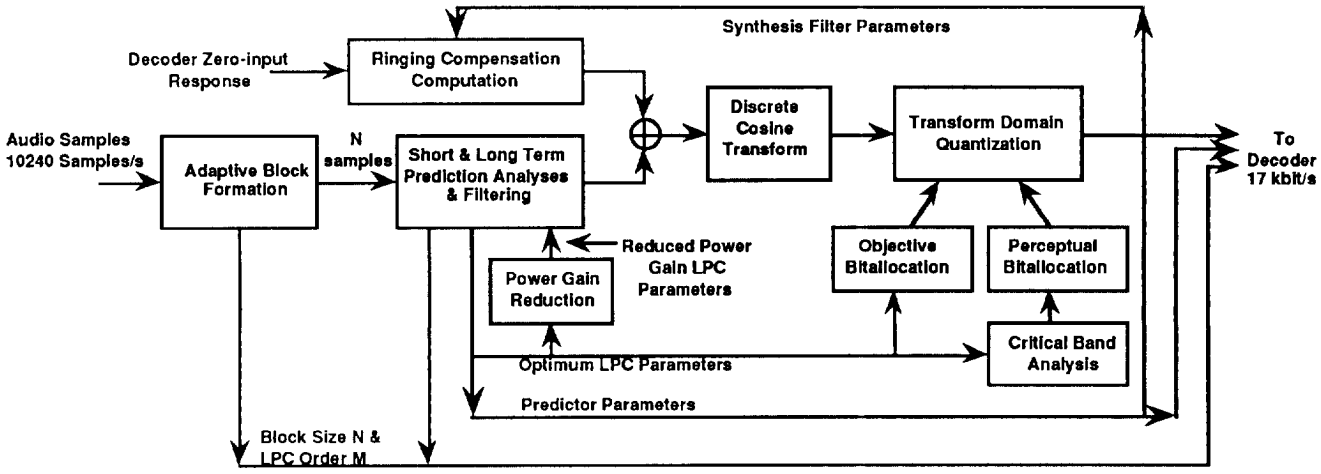

Figure 1. APC-TQ Encoder Block Diagram