Audio Coding Using Sinusoidal Excitation Representation

Wen-Whei Chang, De-Yu Wang, and Li-Wei Wang
Department of Communication Engineering
National Chiao-Tung university
Hsinchu, Taiwan, Republic of China

Abstract— **Most LPC-based audio coders employ simplistic noise-shaping operations to perform psychoacoustic control of quantization noise. In this paper, we report on new approaches to exploiting perceptual masking in the design of adaptive quantization of LPC excitation parameters. Due to its localized spectral sensitivity, sinusoidal excitation representation is preferred to spectrally flat signals for use in excitation modeling. Simulation results indicate that the proposed multisinusoid excited coder can deliver high quality audio reproduction at the rate of 72 kb/s.**

I. Introduction

Since many years there has been considerable interest in high-quality digital audio transmission at lower bit rates. Most perceptual coding systems divide audio spectra into critical bands and then quantize them in accordance with the estimated masking threshold [1]. On the other hand, an LPC-based coder considers audio waveforms to be outputs from an all-pole filter that uses spectrally flat excitation signals (Gaussian white noise for unvoiced signals and multiple impulses for voiced signals). Unfortunately, two observations have been made concerning the inappropriateness of using LPC-based techniques to encode audio signals [2]. First, analysis of experimental data shows that real residual spectra exhibit predominantly pulselike trends, which contrasts sharply with spectrally flat excitation representation. Secondly, most psychoacoustic experiment results are expressed in the frequency domain and hence are not directly applicable for use in conjunction with LPC models.

The strategy applied here is to represent excitation waveforms as a sum of sine waves with arbitrary frequencies, amplitudes, and phases [3]. From the perspective of noise-shaping, this sinusoidal excitation representation provides an ideal framework for incorporating perceptual informations since individual sinusoids can be independently quantized without the leakage of quantization noise from one spectral line to another. This error localization property also helps in developing a dynamic bit-allocation scheme required for perceptually optimal quantization of excitation parameters.

II. Multisinusoid LPC coders

Conventional LPC coders use spectrally flat signals to represent their excitation sources. To better match the peaky residual spectra, we propose to represent excitation waveforms as a sum of sine waves with arbitrary amplitudes, frequencies, and phases. Accordingly, the general form of a multisinusoid excitation model is given by

$$e(n) = \sum_{i=1}^{M} e_i(n) = \sum_{i=1}^{M} r_i cos(w_i nT + \phi_i), \quad 1 \le n \le N, \quad (1)$$

where N is the subframe length, M is the number of sinusoids, and the r_i, w_i and ϕ_i represent amplitude, frequency, and phase, respectively, of the ith sinusoidal component $e_i(n)$. Fig. 1 illustrates the functional block diagram of the proposed MultiSinusoid LPC (MSLPC) encoder. The proposed system performs psychoacoustic control of quantization noise by using a perception-based bit allocation, instead of using a noise-weighting filter for excitation search, as do conventional LPC coders. Two basic types of system parameters can be identified: LPC parameters and excitation parameters. The LPC analysis is performed with autocorrelation method once per frame, whereas excitation parameters are updated once per subframe. In our study, monophonic audio signals with a bandwidth of 15 kHz were sampled at 32 kHz and then segmented into frames of 300 samples long. Each frame was further divided into 6 subfarmes. Letting $h(n)$ denote the impulse response of the synthesis filter, we produce output signals $y(n)$ by taking the convolutional sum

$$y(n) = \sum_{i=1}^{M} [\alpha_i h_{ci}(n) + \beta_i h_{si}(n)], \quad 1 \le n \le N, \quad (2)$$

where $\alpha_i = r_i cos\phi_i$, $\beta_i = -r_i sin\phi_i$, $h_{ci}(n) = cos(w_i nT) * h(n)$, and $h_{si}(n) = sin(w_i nT) * h(n)$.

Accurate identification of excitation parameters can be accomplished by minimizing the squared-error distortion between the original signal $x(n)$ and the output signal $y(n)$. This minimization process resulted in the matrix form of

$$\vec{S} \cdot \vec{g} = \vec{c}, \quad (3)$$

where the entries in $\vec{g}$, $\vec{c}$ and $\vec{S}$ are given as follows for $1 \le j \le 2M$ and $1 \le k \le 2M$, respectively

$$g_j = \begin{cases} \alpha_{(j+1)/2}, & \text{j: odd} \\ \beta_{j/2}, & \text{j: even} \end{cases} \quad (4)$$

This work was supported by the National Science Council, Taiwan, ROC, under Grant No. NSC83-0404-E009-022.

$$c_j = \begin{cases} \vec{x} \cdot \vec{h}^t_{c(j+1)/2}, & \text{j: odd} \\ \vec{x} \cdot \vec{h}^t_{s(j/2)}, & \text{j: even} \end{cases} \quad (5)$$

$$S_{jk} = \begin{cases} \vec{h}_{c(j+1)/2} \cdot \vec{h}^t_{c(k+1)/2}, & \text{j: odd,k: odd} \\ \vec{h}_{s(j/2)} \cdot \vec{h}^t_{s(k/2)}, & \text{j: even,k: even} \\ \vec{h}_{c(j+1)/2} \cdot \vec{h}^t_{s(k/2)}, & \text{j: odd,k: even} \\ \vec{h}_{s(j/2)} \cdot \vec{h}^t_{c(k+1)/2}, & \text{j: even,k: odd.} \end{cases} \quad (6)$$

Using the Cholesky factorization theorem [4], the equation above can be solved more efficiently by decomposing the symmetric matrix $\vec{S}$ into the form of $\vec{G}\vec{G}^t$, where $\vec{G}$ is a lower triangular matrix with non-zero entries as follows:

$$G_{jj} = \sqrt{S_{jj} - \sum_{k=1}^{j-1} G^2_{jk}}, \ 1 \le j \le 2M \quad (7)$$

$$G_{jk} = (S_{jk} - \sum_{l=1}^{k-1} G_{jl}G_{kl})/G_{kk}, \ 1 \le k \le j-1. \quad (8)$$

Proceeding in this way, we can rewrite (3) as follows:

$$\vec{G}\vec{q} = \vec{c} \quad (9)$$

$$\vec{G}^t\vec{g} = \vec{q}, \quad (10)$$

where the entries in $\vec{q}$ are given by

$$q_j = (c_j - \sum_{k=1}^{j-1} G_{jk}q_k)/G_{jj}, \ 1 \le j \le 2M. \quad (11)$$

Using this notation, the least-squared-error distortion is given by

$$E^{(M)}_{min} = E^{(M-1)}_{min} - (q^2_{2M-1} + q^2_{2M}). \quad (12)$$

From inspection of (12), it is evident that the optimum values of the parameters $\{w_i\}$ and $\{r_i, \phi_i\}$ can be independently estimated. As regards the frequencies, a set of L candidates was chosen once per frame by locating the predominant peaks inherent in the associated audio spectrum. Next, only these L candidates were examined to find the M best frequencies needed within each of its 6 constituent subframes. Towards this end, the frequency of the ith component sine wave was taken as the location of the particular candidate, maximizing the term $(q^2_{2i-1}+q^2_{2i})$. Once the frequencies were determined, the optimal values of $\{r_i, \phi_i\}$, which are exclusively embedded in $\vec{g}$, could be found by solving equation (10).

III. Quantization and bit allocation

In this paper, we are more concerned with efficient quantization aspects of LPC parameters and excitation parameters. The class of audio coders discussed here were designed to operate at the rate of 72 kb/s. Using an analysis frame length of 9.375 msec, the total number of bits available per frame is 675, with bits allocated to parameters as listed in Table I. As the table shows, we need to transmit 78 bits per frame as side information regarding the adaptation of bit allocation to time-varying input signal variances.

A. LPC parameters

In this experiment, 10th-order LPC analysis was chosen to characterize the spectral envelope information of incoming sound. Prior to transmission, these LPC parameters were transformed into line spectral frequencies (LSF's) and then quantized using split-vector quantization at 24 bits/frame [5]. More explicitly, we divided the vectors of 10 LSF's into two parts: one consisting of the first 4 LSF's and the other consisting of the remaining 6 LSF's. Each of these two parts was equally allocated 12 bits. We first examined whether LSF parameters could be efficiently quantized using split-vector quantization. The monophonic audio database for these studies consisted of 200 seconds of audio signals recorded from various musical instruments. The first 170 seconds of music was used for training, and the last 30 seconds of music was used for testing. The performance was evaluated in terms of spectral distortion (SD), defined as the root mean square difference between the original LPC log-power spectrum and its quantized version. An average SD of 1 dB is usually accepted as the difference margin for spectral transparency. Since no SD scores exceeded 1 dB for any of our test samples, we can conclude that a split-vector quantizer can represent ten LPC parameters at 24 bits per frame with transparent quality.

B. Excitation parameters

The excitation parameters discussed here consist of the frequencies, amplitudes, and phases of the component sine waves. As regards the frequencies, a set of 13 spectral peaks per frame were first located as candidates and then examined to find the 7 best frequencies needed within each of its 6 constituent subframes. Since the frequency range was resolved in a 1024-point DFT, a direct approach to representing each candidate position required the use of 9 bits. To save bit quota, we encoded the first candidate as an absolute location in the frame and the remaining candidates as differences from the previous one. To elaborate further, these 13 candidates were differentially encoded with 78 bits in accordance with the bit allocation (5,6,6,6,6,6,6,6,6,6,6,6,7). Next, we employed an enumerative source coding technique [6] to encode the 7 best frequencies once per subframe. Since the number of different possibilities involved in choosing 7 out of 13 candidates is given by C^{13}_7, the minimum number of bits required to encode all possible patterns within a subframe is 11.

To quantize the amplitudes an adaptive quantizer whose levels were adjusted to the maximum absolute value within a frame was used. This maximum absolute value, denoted by r_{max}, was logrithmically encoded in 9 bits. The individual amplitudes were then scaled and uniformly quantized using varying degrees of bit resolution. The aim was to obtain a larger margin between the coder generated noise level and the audibility threshold of such artifacts. Following the work described in [7], we first implemented a perceptual model to obtain the input parameters (mask-to-noise ratios) required to optimize the bit-rate adjustment procedure. The calculation started with a precise spectral anal-

ysis on 1024 windowed audio samples to generate its magnitude spectrum. The spectral lines were then examined to discriminate between tonelike and noiselike maskers by taking the spectral flatness measure as an indicator of tonality. Using rules known from psychoacoustics, the spread Bark spectrum was then calculated dependent on frequency position, loudness level, and the nature of tonality. Finally, we obtained a vector of 24 masking thresholds, denoted by $\{mask(b), b = 1, 2, \ldots, 24\}$, from the spread Bark spectrum and from the absolute threshold in quiet.

Constrained to producing a constant bit-rate for each frame, we proposed a dynamic bit allocation routine based on the MNR (mask-to-noise ratio) perceptual measure, which is defined as the ratio of the estimated masking threshold to actual coding noise. The primary goal was to minimize the total mask-to-noise ratio over each subframe by increasing the quantizer resolution for perceptually more important sinusoids until the number of bits available was exhausted. Let us assume that σ_i^2 is the variance of the i-th filtered sinusoidal component, denoted by $s_i(n) = e_i(n) * h(n)$, and let $\sigma_{q_i}^2$ denote the quantization noise variance associated with R_i-bit uniform quantization of the signal $s_i(n)$. The proposed bit allocation routine is an iterative procedure, where in each iteration the following steps proceed until all 35 bits have been allocated in coding the amplitudes.

(1) Calculate the error variances of all the sinusoids, $1 \leq i \leq 7$,

$$\sigma_{q_i}^2 = \epsilon \sigma_i^2 / 2^{2R_i} \tag{13}$$

where ϵ is the corresponding quantizer performance factor.

(2) Calculate the MNR of all the sinusoids, $1 \leq i \leq 7$,

$$MNR(i) = mask(b) - 10log_{10}\sigma_{q_i}^2, wl_b \leq w_i \leq wh_b, \tag{14}$$

where wl_b and wh_b denote, respectively, the lower and the upper boundaries of the bth critical band.

(3) Assign one additional bit to the particular sinusoid with the minimum MNR.

Once the final bit allocation was determined, the individual amplitudes were then uniformly quantized in the range of $[0, r_{max}]$ with different quantizer resolutions. As regards the phases, each sine wave was equally allocated 5 bits and uniformly quantized in the range of 0 to 2π.

IV. Experimental results

Computer simulations were conducted to examine the suitability of MNR-adapted bit allocation for use with the multisinusoid excited LPC coder. The monophonic audio database for these studies consisted of electrified instrumental music, an oboe plus a piano, and an orchestra. Each music signal is 10 seconds in duration and sampled at 32 kHz. Though better performance can always be obtained by increasing the number of sine waves, the paramters $L = 13$ and $M = 7$ were empirically chosen as the best compromise between coding gain and implementational complexity.

Table II shows the comparative performance results for 72 kb/s audio coding in conjunction with a multipulse excited model (MPLPC) and a multisinusoid excited model (MSLPC). The performance is evaluated in terms of SNR, segmental SNR (SNRSEG), and Bark spectral distortion (BSD). The BSD measure has been shown to correlate more closely with the results of human preference tests than those obtained by other conventional objective measures [8]. As the table shows, the MSLPC coder yielded substantial improvement over the MPLPC coder for all test samples. Our informal listening tests also confirmed the superior quality of the MSLPC output. Among the reasons for success, we found that sinusoidal excitation representation can more closely match the intrinsic natures of actual residual spectra. Compared to the MPLPC case, an MSLPC coder has one additional advantage of using MNR-adapted bit allocation to increase quantizer resolution for psychoacoustically important sinusoidal components.

V. Conclusions

In this paper, we first emphasized the importance of matching LPC excitation sources to the pulselike natures of residual spectra. This was done by using a sum of sine waves to approximate LPC excitation waveforms rather than using spectrally flat excitation signals. Furthermore, it was found that sinusoidal excitation representation provides an ideal framework for incorporating masking thresholds in the design of noise spectral shaping. Experimental results concluded that the use of sinusoidal excitation representation combined with a perception-based quantization allows the implementation of an LPC audio coder that delivers high quality at the rate of 72 kb/s.

References

[1] J. D. Johnson, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, pp. 314-323, Feb. 1988.

[2] W. W. Chang and C. T. Wang, "A masking-threshold-adapted weighting filter for excitation search," *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 124-132, March 1996.

[3] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal model," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744-754, August 1986.

[4] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, 1989.

[5] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, pp. 3-14, Jan. 1993.

[6] T. M. Cover, "Enumerative source coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 73-77, Jan. 1973.

[7] ISO/IEC Int. Std. IS 11172-3, "Information technology-Coding of moving pictures and associated audio for digital storage media up to about 1.5 mbits/s," Part 3: Audio.

[8] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 819-829, June 1992.

Table 1: Bit allocation for MSLPC coders at 72 kb/s

Amplitudes	35 × 6
Phases	35 × 6
Frequencies	11 × 6
Frequency Candidates	78
Maximum Amplitude	9
LPC Parameters	24
Bit Allocation Information	13 × 6
Total bits per frame	675

Table 2: SNR/SNRSEG/BSD performances of MPLPC and MSLPC coders at 72 kb/s

Coder / Music	MPLPC	MSLPC
Electric Instrument	24.52 / 24.71 / 21.00	24.38 / 26.61 / 20.73
Oboe + Piano	23.89 / 24.47 / 72.14	24.55 / 27.92 / 6.98
Orchestra	24.23 / 24.12 / 119.4	24.82 / 26.43 / 149.10

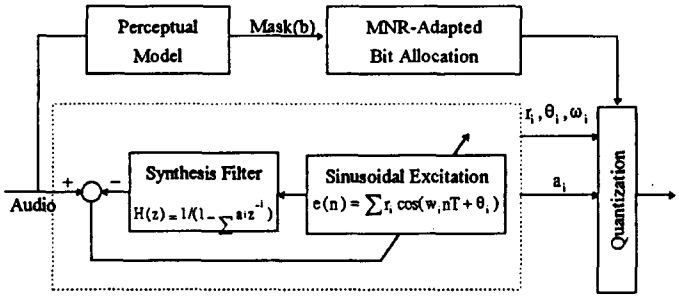

Fig. 1. Block diagram of the multisinusoid-excited LPC encoder