SPEECH CODING WITH AN ANALYSIS-BY-SYNTHESIS SINUSOIDAL MODEL

Çağrı Özgenç Etemoğlu, Vladimir Cuperman and Allen Gersho

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106
E-mail:[cagri, vladimir, gersho]@scl.ece.ucsb.edu

ABSTRACT

We introduce a general and powerful approach to sinusoidal modeling of speech wherein a closed-loop Analysis-by-Synthesis (AbS) technique sequentially extracts the parameters for each sinusoidal component. Low bit-rate speech coding is achieved by efficiently constraining the allowed frequencies of sinusoidal components into sets of frequency intervals or bins. In conjunction with the closed-loop analysis, the constrained frequency regions allow us to efficiently vector quantize the frequency information in each frame. In voiced frames, two sets of frequency vectors are generated: one for harmonically related components and the other for non-harmonically related components of the voiced segment. In transition frames, a vector of nonuniformly spaced frequencies is selected from a frequency codebook using frequency bin vector quantization (FBVQ) to represent the frequency domain information. The effectiveness of the coding scheme is enhanced by exploiting the critical band concept of auditory perception in defining the frequency bins. In transition segments, the sinusoidal phases are modeled and coded. Subjective tests with a partially quantized model indicate that, for a target rate of 4 kbps, the coder quality exceeds that of the G.729 standard at 8 kbps.

1. INTRODUCTION

It is well-known that, *code-excited linear predictive* (CELP) coding is able to achieve toll or nearly toll quality speech at rates above 5 kbps. However, below 5 kbps, the speech quality of CELP coders degrades due to its inability to accurately match the speech waveform with the inadequate number of excitation bits available for the frame. On the other hand, parametric coders such as the *sinusoidal-transform coder* (STC) [1], the *waveform-interpolative* (WI) coder [2], and the *multiband-excitation* (MBE) coder [3] can produce good quality speech at rates as low as 2 kbps. These coders do not achieve toll quality and lack robustness to different speakers. We believe these deficiencies are partly caused by the open-loop character of the sinusoidal coders and partly by their inability to model transition segments such as voicing onsets and plosives.

This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, Cisco Systems, Inc., Dialogic Corp., Fujitsu Laboratories of America, Inc., General Electric Co., Hughes Network Systems, Intel Corp., Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Qualcomm, Inc., Rockwell International Corp., Panasonic Technologies, Inc., and Texas Instruments, Inc.

In this paper, we introduce an Analysis by Synthesis (AbS) sinusoidal modeling technique for low bit-rate speech coding wherein the parameters for each sinusoidal component are sequentially extracted by a closed-loop analysis. The sinusoidal modeling of the speech residual is performed within the general framework of matching pursuits [4, 5] with a dictionary of sinusoids. The frequency range is restricted to sets of frequency intervals or *bins*, which in conjunction with the closed-loop analysis allow us to map the frequencies of the sinusoids into a frequency vector that is efficiently quantized. In voiced frames, two sets of frequency vectors are generated: one of them represents harmonically related and the other one non-harmonically related components of the voiced segment. This approach eliminates the need for voicing information that is difficult to estimate correctly and to quantize at low bit rates. In transition frames, a vector of nonuniformly spaced frequencies is selected from a frequency codebook using frequency bin vector quantization (FBVQ) to represent the frequency domain information. Our use of FBVQ with closed-loop searching combined with modeling and coding of the perceptually important phase information together contribute to a significant improvement of speech quality in transition frames. Subjective tests indicate that a partially quantized model with a target rate of 4 kbps has quality exceeding the G.729 standard at 8 kbps.

2. ANALYSIS/SYNTHESIS

2.1. Synthesis Model

In our model, each frame of the linear prediction (LP) residual is represented as a sum of sinusoids which are weighted by a magnitude envelope $\sigma^k[n]$. Thus, for the k^{th} frame, we have

$$\hat{s}^k[n] = \sigma^k[n] \sum_{i=1}^{M} A_i^k \cos(\omega_i^k n + \phi_i^k) \tag{1}$$

The parameter set for each frame consists of an amplitude vector $\boldsymbol{A} = \{A_i\}$, a frequency vector $\boldsymbol{\omega} = \{\omega_i\}$, and a phase vector $\boldsymbol{\phi} = \{\phi_i\}$. The synthesized frames are combined by using overlap-add to obtain the reconstructed LP residual, $\hat{s}[n]$

$$\hat{s}[n] = \sum_k W_s[n - kN_s]\hat{s}^k[n] \tag{2}$$

where N_s is the synthesis frame size. The synthesis window obeys the constraint

$$\sum_k W_s[n - kN_s] = 1 \tag{3}$$

2.2. Analysis with Matching Pursuits

To effectively represent the LP residual as a sum of sinusoids, we adopt the general approach of matching pursuits [5]. This is an iterative algorithm, which represents a given signal in terms of a linear combination of a set of waveforms, selected sequentially from a redundant dictionary whose size is generally much larger than the number of terms needed for an adequate representation. In our case, the dictionary $\mathcal{D}$ is a set of magnitude envelope weighted cosine waveforms as described in Section 2.1. The frequencies $\{\omega_j\}$ of the cosine waveforms forming the dictionary are defined by using a fine grid of L points ($L \gg M$) covering the spectral range of interest and given by $\omega_j = j\pi/(L-1)$ for $j = 0, 1, \ldots, L-1$. The frequencies, amplitudes, and phases for each term in the representation are parameters to be determined by the modeling algorithm. The sum of sinusoids is weighted by a magnitude envelope $\sigma[n]$, to track speech energy variations across the frame. Later we describe how this envelope is obtained and efficiently quantized. In each iteration, a new sinusoidal term is added to the model, then the modeling error waveform (error residual) is formed. The parameters for each sinusoid is optimized to minimize a weighted measure of the error residual energy. Thus, the error residual after m iterations, $r_m[n]$, is given by

$$\begin{aligned} r_m[n] &= r_{m-1}[n] - \sigma[n] A_m \cos(\omega_m n + \phi_m) \qquad (4) \\ &= s[n] - \sigma[n] \sum_{i=1}^{m} A_i \cos(\omega_i n + \phi_i) \end{aligned}$$

where for simplicity the frame index k is omitted.

At the m^{th} iteration, the algorithm will search for the frequency point ω_m, which together with its optimal amplitude A_m and optimal phase ϕ_m minimizes the weighted energy E_m of the error residual given by

$$E_m = \sum_{n \in \mathcal{N}} W_a[n] \{ r_{m-1}[n] - \sigma[n] A_m \cos(\omega_m n + \phi_m) \}^2 \quad (5)$$

where $\mathcal{N}$ denotes the time span of the current analysis frame. The analysis window $W_a[n]$ serves as a weighting in equation 5 and enhances the representation of the region in which $\hat{s}_k[n]$ has the dominant contribution to $\hat{s}[n]$.

While this algorithm is able to synthesize high quality speech, it has two major drawbacks. First, the computational complexity is very high, since at each iteration it eliminates only one frequency point and searches through essentially the entire grid of finely-spaced frequencies. Second, the resulting set of frequencies, representing the frame, are irregularly spaced and therefore are difficult to quantize at low bit rates.

These two problems motivated us to develop a novel *dynamic dictionary* matching pursuits algorithm based on a *frequency bin model* for structuring and simplifying the allowed set of sinusoidal component frequencies in the dictionary. We refer to this set of frequencies as the *frequency space* of the dictionary.

2.3. Analysis with Dynamic Dictionary Matching Pursuits using a Frequency Bin Model

The dynamic dictionary matching pursuits is a modified matching pursuits algorithm, in which the dictionary is updated at each iteration by removing a group of dictionary elements. The complexity of this algorithm is substantially less than that of the conventional matching pursuits algorithm since the size of the dictionary gradually decreases with successive iterations.

The frequency bin structure represents the frequency space of allowed cosine waveforms as a set of non-overlapping frequency intervals or *bins* where each bin consists of the set of frequency grid points contained in that interval. Since only one frequency within a given bin will be used in the decoder's synthesis procedure, the width of each bin is chosen as large as possible while satisfying the rule that the perceptual difference between the center frequency and any other frequency point in the bin should be insignificant when using the model in equation 1. With this requirement the widths of the bins must increase with increasing frequency, since the human auditory system's frequency resolution decreases as the frequency increases. This rule further guarantees that any frequency point in a bin can be quantized to that bin's center frequency without sacrificing perceptual information.

The frequency bin model combines with the dynamic dictionary matching pursuits as follows. At each iteration, the analysis procedure will choose the best matching frequency point from the frequency space (determined by the current set of bins); then the dictionary is updated by removing the entire bin corresponding to that frequency point. After all the bins are eliminated, the analysis will stop. Therefore, the number of iterations will be equal to the number of bins in the frequency space that forms the initial dictionary. This search process determines a set of sinusoids whose frequencies are still unquantized. For encoding, these frequency points are quantized to the center frequencies of their respective bins.

Specifically, for a given magnitude envelope $\sigma[n]$, at the m^{th} iteration, given the current dictionary $\mathcal{D}_{m-1}$, and the current residual $r_{m-1}[n]$, we search the frequency space for the frequency point ω_m that minimizes equation 5. Then we update the residual by using equation 4 and finally update the dictionary, $\mathcal{D}_{m-1} \rightarrow \mathcal{D}_m$, by removing the bin in which ω_m resides from $\mathcal{D}_{m-1}$.

The analysis procedure described in this section assumes a given dictionary with an associated set of frequency bins. The next section applies this analysis to search a family of dictionaries, represented by a vector quantization codebook.

3. FREQUENCY BIN VECTOR QUANTIZATION

3.1. Quantization Method

To efficiently quantize the set of frequencies needed for the sinusoidal representation of the LP residual in transition frames, we introduce Frequency Bin Vector Quantization (FBVQ). In FBVQ, encoding is based on a pair of codebooks, a frequency codebook $\mathcal{C}_\mathcal{F}$ with elements $\boldsymbol{c_j}$ and a bin width codebook $\mathcal{B}$ with elements $\boldsymbol{\Delta_j}$, where the vector $\boldsymbol{c_j}$ is an ordered set of M frequency values and $\boldsymbol{\Delta_j}$ is an ordered set of corresponding bin widths. From the j^{th} pair of codevectors we can generate a dictionary of sinusoids $\mathcal{D}_0^j$, whose frequency space consists of all frequency grid points in the bins that are centered at the elements of $\boldsymbol{c_j}$ and have widths given by the bin widths vector $\boldsymbol{\Delta_j}$.

Figure 1 shows a block diagram of the AbS analysis procedure based on FBVQ. In the figure M is the dimension of the frequency codevector. At the m^{th} iteration corresponding to the

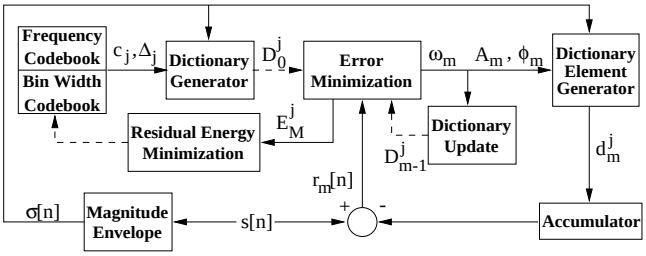

Figure 1: Block diagram of FBVQ

j^{th} frequency and bin width codevectors, the best matching dictionary element is denoted by d_m^j. Note that the vector dimension M is also the number of iterations, therefore E_M^j denotes the final residual energy corresponding to the dictionary generated by the j^{th} pair of codevectors, $\boldsymbol{c_j}$ and $\boldsymbol{\Delta_j}$. The magnitude envelope $\sigma[n]$ reduces the effect of energy variations on the estimation of the sinusoidal parameters. Later one possible approach to obtain and quantize magnitude envelope is described. The analysis procedure in conjunction with FBVQ finds the index of the codebook entry that best matches the signal perceptually, and calculates the corresponding amplitude and phase vectors. FBVQ will use the analysis to select the codebook index, whose corresponding dictionary yields the minimum residual energy E_M^j. In this context, the analysis will act as a method for computing the metric for Nearest Neighbor (NN) Condition of FBVQ.

3.2. Codebook Design Issues

An effective design is needed for the frequency and bin width codebooks. A possible approach for selecting the frequency codevector $\boldsymbol{c_j}$ is to uniformly sample the conventional frequency scale, but this would not account for the non-uniform frequency resolution of the human auditory system. A better approach is to sample the equivalent rectangular bandwidth (ERB) rate scale [6] uniformly, since ERB-rate scale, like human auditory system, has a decreasing frequency resolution with increasing frequency.

The second design issue is the choice of the bin widths vector $\boldsymbol{\Delta_j}$. At low bit rates, we would like to model the input signal by using as few cosine waveforms as possible, since the number of parameters to encode is proportional to the number of cosine waveforms. Experimental evidence shows that we can synthesize good quality voiced frames by using a number of cosines equal to the number of pitch harmonics and good quality transition frames by 20-30 cosines. Once the number of cosines is fixed, the number of bins will be the same, since the analysis generates a single cosine corresponding to each bin. Given the number of bins for a given type of frame, the size of the dictionary will be proportional to the bin widths. In this case $\boldsymbol{\Delta_j}$ determines the trade-off between the modeling error caused by the reduction of the frequency space into bins and the quantization error caused by mapping the frequency point selected from each bin to that bin's center frequency. If we choose $\{\Delta_{jk}\}$'s too small, the quantization error will be negligible which means the reconstructed signal in the analyzer will be very similar to the one in the synthesizer, but the resulting small dictionary may not accurately model the input waveform. On the other hand, choosing $\{\Delta_{jk}\}$'s too large will lead to a large dictionary which can model the input waveform well, so the reconstructed signal in the analyzer will be of high quality, but the reconstructed signal in the synthesizer will suffer from large quantization errors. A good trade-off can be obtained by increasing Δ_{jk} up to the point where the quantization error is still perceptually insignificant. In general Δ_{jk} will increase with frequency, since the human auditory system is more tolerable to quantization errors at higher frequencies.

4. VOICED, TRANSITION, AND UNVOICED ANALYSIS/SYNTHESIS

4.1. Voiced/Unvoiced Analysis and Synthesis

To efficiently model speech, the phonetic character of individual frames should be considered. For voiced frames, we use two frequency vectors to capture the frequency domain information. One of them is composed of harmonically related frequencies $\boldsymbol{\omega_h}$ and represents the periodic part; the other one is composed of non-harmonically related frequencies $\boldsymbol{\omega_{nh}}$ and represents the aperiodic part of the voiced segment. The elements of $\boldsymbol{\omega_h}$ are multiples of $\omega_o = 2\pi/p_o$, where p_o is the pitch period. The elements of $\boldsymbol{\omega_{nh}}$ are obtained by uniformly sampling the portion of the ERB-rate scale above 1 kHz. Typical voiced frames do not have large energy variations across the frame, therefore the magnitude envelope is set to $\sigma[n] = 1$.

The voiced residual is modeled as a sum of the harmonic and the non-harmonic models, which have $\mathcal{D}_h$ and $\mathcal{D}_{nh}$ respectively as their dictionaries. The initial dictionary $\mathcal{D}_0 = \mathcal{D}_h$ corresponds to harmonic analysis. Harmonic analysis generates the frequency $\boldsymbol{\omega_p}$, amplitude $\boldsymbol{A_p}$, and phase $\boldsymbol{\phi_p}$ vectors representing the periodic part. In harmonic analysis, the frequency space of dictionary $\mathcal{D}_h$ consists of bins which are centered at the pitch harmonics. The frequency points generated during analysis do not have to be exact multiples of ω_o, enabling harmonic analysis to capture periodic components even in frames that have changing pitch period. Therefore the leakage from periodic to aperiodic part will be reduced, resulting in an error residual of negligible periodicity, which is suitable as an input for non-harmonic analysis. After K iterations, (K being the number of pitch harmonics, $K = \lfloor p_o/2 \rfloor$) the dictionary is set to $\mathcal{D}_K = \mathcal{D}_{nh}$ for non-harmonic analysis. Non-harmonic analysis works on the error residual generated by the harmonic analysis. It generates the frequency $\boldsymbol{\omega_{ap}}$, amplitude $\boldsymbol{A_{ap}}$, and phase $\boldsymbol{\phi_{ap}}$ vectors representing the aperiodic part. The frequency space of dictionary $\mathcal{D}_{nh}$ consists of bins which have the elements of $\boldsymbol{\omega_{nh}}$ as their center frequencies.

For the synthesis of the periodic part, a cubic phase model [1] for the zero dispersion phase is used with the frequency $\boldsymbol{\omega_h}$, and the quantized amplitude $\hat{\boldsymbol{A}}_{\boldsymbol{p}}$ vectors. The linear phase is formed by keeping track of successive onset times generated by the succession of pitch periods that are available at the decoder [7]. The aperiodic part is synthesized by applying uniformly distributed random phases to cosines having the frequency $\boldsymbol{\omega_{nh}}$, and the quantized amplitude $\hat{\boldsymbol{A}}_{\boldsymbol{ap}}$ vectors.

Unvoiced analysis is the same as harmonic analysis with bins located at multiples of 100 Hz, with the exception that a magnitude envelope is used as in the case of transition frames. The synthesizer applies uniformly distributed random phase to each frequency component.

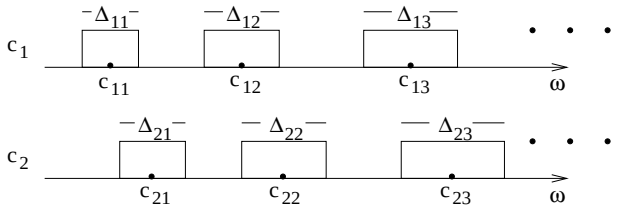

Figure 2: Sample codevectors and their bin structure

4.2. Transition Analysis and Synthesis

Transition frames are represented by a frequency vector $\hat{\boldsymbol{\omega}}_t$ selected from a frequency codebook $\mathcal{C}_{\mathcal{F}}$ by using FBVQ, and its corresponding amplitude $\boldsymbol{A}_t$ and phase $\boldsymbol{\phi}_t$ vectors. The codebook has 8 codevectors which are obtained by uniformly sampling the ERB-rate scale. Two codevectors and their corresponding frequency bins structure are illustrated in Figure 2. The magnitude envelope $\sigma[n]$ is formed by linearly interpolating the magnitude vector, $\boldsymbol{\sigma} = \{\sigma_i\}$

$$\sigma_i = \sqrt{\sum_j a_j s^2[l_p + iD - j]} \quad i = 0, 1, \ldots \tag{6}$$

where D is the downsampling factor, and l_p is the peak energy position used to align downsampling operation with the largest energy peak. This alignment alleviates the smearing of magnitude envelope caused by downsampling and linear interpolation.

The quantized magnitude envelope, $\hat{\sigma}[n]$ is formed by linearly interpolating the quantized magnitude vector $\hat{\boldsymbol{\sigma}}$, whose elements have time locations aligned with l_p. The peak energy position estimate l_p is scalar quantized using 4 bits. The magnitude vector $\boldsymbol{\sigma}$ is quantized using a gain-shape decomposition. Five bits are used to quantize the shape of the magnitude vector.

At the decoder, a cubic phase model with the frequency $\hat{\boldsymbol{\omega}}_t$, the quantized phase $\hat{\boldsymbol{\phi}}_t$, and the quantized amplitude $\hat{\boldsymbol{A}}_t$ vectors is used to synthesize transition frames.

5. PHASE QUANTIZATION

Two different procedures are used to quantize transition phase vector $\boldsymbol{\phi}_t$ and the procedure having the lower closed-loop distortion is selected. The first procedure decomposes the transition phase into a linear phase and dispersion phase as follows:

$$\boldsymbol{\phi}_t = \boldsymbol{\omega}_t n_0 + \boldsymbol{\psi}_t \tag{7}$$

where n_0 represents the linear component and $\boldsymbol{\psi}_t$ the dispersion phase. The dispersion phase vector $\boldsymbol{\psi}_t$ is quantized using a mean-shape decomposition. The mean and shape phase components are quantized jointly by using a uniform scalar quantizer for the mean and a codebook (generated as uniformly distributed random vectors of narrow dynamic range) for the shape. The second procedure uses a codebook of uniformly distributed random phases for quantization. The transition phase vector is quantized using 10 bits, where one of the bits is used to distinguish between procedures.

6. BIT ALLOCATION

The bit allocation for each type of speech frame is given in Table 1, however in our current simulation, the LPC and amplitude parameters have not yet been quantized. The frame size is 20 ms.

Parameter	Transition	Voiced	Unvoiced
LPC	18	20	18
Amplitude	15	43	39
Phase	$2 \times 10 = 20$	0	0
Frequency	$2 \times 3 = 6$	0	0
Pitch	0	$2 \times 7 = 14$	0
Envelope	$2 \times 5 = 10$	0	$2 \times 10 = 20$
Peak Eng. Pos.	$2 \times 4 = 8$	0	0
Classifier	3	3	3
Total	80	80	80
Bit-rate	4kbps	4kbps	4kbps

Table 1: Bit allocation

7. SUBJECTIVE RESULTS

We have conducted a preference listening test to compare the subjective performance of the proposed AbS coder with the G.729 standard. The test data included 16 MIRS speech sentences, 8 from female speakers and 8 from male speakers. Twelve listeners participated in the test. The test results presented in Table 2, indicate that the subjective quality of the proposed partially quantized coder exceeds that of G.729 at 8kbps.

Speakers	AbS coder	G.729	Same
Female	42.71%	32.29%	25.00%
Male	52.08%	19.79%	28.13%
Total	47.40%	26.04%	26.56%

Table 2: Preference test results

8. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, August 1986.

[2] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 386–399, October 1993.

[3] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1223–1235, August 1988.

[4] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 389–406, September 1997.

[5] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3397–3415, December 1993.

[6] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 115–132, January 1994.

[7] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), ch. 4, pp. 144–145, Elsevier, 1995.