TECHNIQUES FOR IMPROVING SINUSOIDAL TRANSFORM VOCODERS

Wen-Whei Chang and De-Yu Wang

Department of Communication Engineering National Chiao Tung University
Hsinchu, Taiwan, Republic of China
e-mail: wwchang@cc.nctu.edu.tw

ABSTRACT

This paper presents quality enhancement of sinusoidal transform coders (STC) via the development of new parametric models. First explored are the benefits of Bark spectrum for use in the design of perceptual coding of the sine-wave amplitudes. According to our results, the proposed approach provides a uniform perceptual fit across the spectrum. To enhance the accuracy of phase representation, noncausal all-pole modeling of the vocal system is also discussed. Experimental results indicate that the use of new parametric models allows the STC to improve the phase accuracy as well as the synthetic speech quality.

1. INTRODUCTION

Recent developments in STC technology have made possible synthetic speech of good quality at very low data rates [1]. STC attempts to model speech waveforms as the sum of sinusoids whose frequencies, amplitudes, and phases are chosen to make the reconstruction a best fit to the original speech. One way to encode these parameters at low rates is to exploit a minimum-phase harmonic sine-wave speech model, in which the sine-wave frequencies are harmonically related, and amplitudes and phases are represented in terms of cepstral coefficients. The basic problem with cepstral representation is that the modeling accuracy tends to be uniform across all frequencies and cannot precisely describe the ear's nonlinear responses to frequency selectivity and subjective loudness. This shortage can be partially alleviated by warping the frequency axis to give more prominence to the perceptually more important frequencies [1]. The present work attempts to capitalize more fully on psychoacoustic knowledge and then develop an amplitude coder based on the Bark spectrum [2], instead of those based on cepstral representation.

One major advantage of cepstral representation is the possible elimination of the need to code the phase information, by observing that the log magnitude and phase of a minimum phase system satisfy a Hilbert transform relationship [3]. Recent studies, however, indicate the inadequacy of the minimum-phase assumption for modeling voiced speech due to the anticausal nature of the glottal excitation [4,5]. Recognizing this, several refinements of the minimum-phase model have been developed that improve the phase accuracy by using either a Rosenberg pulse model or an all-pass filter [4]. Unlike the current STC, it is proposed herein that the vocal system be modelled by a noncausal filter of all-pole type. The motivation for this representation is given in two ways. First, it has been shown that noncausal all-pole filters are more appropriate for modeling the vocal system because they take into account the maximum-phase poles of differentiated glottal pulses [5]. Second, the minimum-phase assumption is more applicable to versions of STC which codes the sine-wave amplitudes using cepstral coefficients rather than those using the Bark spectrum. This is because that only for a minimum-phase system, its phase response can be explicitly identified by applying a Hilbert transform to the cepstral envelope of sine-wave amplitudes. In contrast, the noncausal all-pole approach proposed ferein applies to both representations.

This work was supported by National Science Council, Taiwan, ROC, under Grant NSC86-2221-E009-026.

2. THE HARMONIC SINE-WAVE MODEL

A promising approach to the parameter quantization problem lies in the observation that the voiced speech, when perfectly periodic, can be represented by harmonic components of its Fourier series decomposition. In this case, the general form of a harmonic sine-wave model can be expressed as:

$$\hat{x}(n) = \sum_{l=1}^{L} A_l cos(nlw_0 + \theta_l) \qquad (1)$$

where L denotes the number of sinusoids, w_0 represents the pitch frequency, A_l and θ_l are the amplitude and phase of the l-th sinusoidal component. Because of the time-varying nature of the parameters, both birth-death frequency tracking and cubic interpolation phase unwrapping techniques must be introduced to ensure a smooth evolution from frame to frame.

A low-rate representation is achievable by fitting a set of cepstral coefficients to an envelope of the measured sine-wave amplitudes [1]. For the system with transfer function $H(z)$, the cepstrum is defined to be the sequence of coefficients in the power series representation of its log magnitude, i.e.,

$$c_m = \frac{1}{\pi} \int_0^{\pi} log|H(w)|cos(mw)dw, \; 0 \leq m \leq M-1. \qquad (2)$$

The main attraction of cepstral representation arises from the fact that it exploits the minimum-phase model, where

the log magnitude and phase of the system function can be uniquely related in terms of the Hilbert transform [3].

With this exploitation, additional economies in coding the phase information can be obtained by explicitly identifying the phase components due to the excitation and the vocal tract. The first step is to employ a mixed excitation model in which below the voicing-adaptive transition frequency the excitation phases are made linear and above the transition they are made random on $[-\pi, \pi]$. When combined with the minimum-phase component derived from the cepstrum, it was shown [1] that synthetic speech of good quality can be obtained without the need to code the phase information.

3. PERCEPTUAL CODING OF SINE-WAVE AMPLITUDES

Perceptual coding is intended not merely for using statistical correlation to remove waveform redundancies, but also for eliminating the perceptual irrelevancy by applying psychoacoustic measures. The main drawback of using cepstral analysis to obtain the smoothed envelope of the sine wave amplitudes is that it leads to a uniform fit across the whole frequency range. This is inconsistent with the fact that the ear is less sensitive to details in the sine-wave amplitudes at higher frequencies than at lower ones. This inconsistency can be alleviated to some extent by warping the amplitude envelope following the perceptually based mel scale before computing the cepstral coefficients [1]. Though spectral warping conceptually satisfies its ability to simulate nonlinear frequency resolution, its suitability to represent perceived loudness is limited. This suggests that further improvement can be achieved through a more precise exploitation of psychoacoustic knowledge. To advance with this, we propose to implement an amplitude coder by using the Bark spectrum [2] rather than using the cepstrum, as do current STC coders [1].

The advantage of the Bark spectrum over the cepstrum is that it more closely emulates known features of human hearing. The calculation of Bark spectrum involves the Hertz-to-Bark transformation, critical-band filtering, equal-loudness preemphasis, and subjective-loudness conversion. In correspondence with the warping function $b = R(f)$, we first derive the critical-band density $X(b)$ by substituting the frequency variable f in the power spectrum $X(f)$ with the Bark scale b. Next, we performed critical-band filtering to determine the excitation pattern $D(b)$ by taking the convolution of $X(b)$ with the basilar-membrane spreading function $F(b)$. Notably the convolution with the relatively broad spreading function significantly reduces the spectral resolution of $D(b)$. This feature allows for down-sampling of the excitation pattern at one-Bark intervals. It typically suffices to use 14 spectral samples of $D(b)$ to cover the 3.4 kHz speech bandwidth. Finally, phon-to-sone conversion is needed to compensate for the difference between the loudness level and the subjective loudness scale. The resulting Bark spectrum $B_X(b)$, which reflects the ear's nonlinear transformation of frequency and loudness, yields a measure in terms of which perceptual information can be more precisely incorporated in the coder design.

While Bark spectral analysis is a necessary first step in developing amplitude coding, there remains the problem of inverse processing in the hope that sine-wave amplitudes can be recovered from the received version of Bark spectrum at the decoder side. This task can be aided by taking advantage of the harmonic modeling assumption, in which the sine-wave amplitudes are replaced by the harmonic samples of the spectral envelope. The strategies for estimating these harmonic amplitudes maybe divided into two steps. First, Bark spectrum is inversely processed to obtain the excitation pattern following the sone-to-phon conversion and equal-loudness deemphasis. The next problem to address is the association of the resulting excitation pattern with harmonic amplitudes. In this respect, it is more convenient to describe the manipulations in terms of matrix algebra. Consider the vectors $\vec{D}$ and $\vec{X}$ representing the excitation pattern and N-point discrete Fourier transform (DFT) of incoming sound. For ease of notation, the frequency corresponding to the j-th DFT coefficient is referred to as $f_j = jf_s/N$, where f_s denotes the sampling rate. The contribution due to the spreading function can be summarized in a matrix $C = [c_{i,j}]$, where the entry $c_{i,j}$ takes the value of $F[R(f_j) - i]$ if the frequency f_j lies within the i-th critical band, and takes the value of 0 elsewhere. With these descriptions, the calculation of excitation pattern $\vec{D}$ can then be formulated as applying the matrix C on $\vec{X}$, i.e.,

$$\vec{D} = C \cdot \vec{X}. \tag{3}$$

Note that the harmonic amplitudes, which are exclusively embedded in the unknown vector $\vec{X}$, are the parameters to be estimated. Unfortunately, however, a unique solution does not exist due to the fact that the number of equations is less than the number of unknowns. To illustrate a typical low-pitched speaker can have as many as 80 harmonic samples in a 4-kHz speech bandwidth, compared to the dimension of 14 in $\vec{D}$. In order to pursue a unique solution, we assign an equal amplitude to those harmonics belonging to the same critical band.

4. NONCAUSAL ALL-POLE MODELING OF VOCAL SYSTEM

At low rates, more properties of the speech production mechanism need to be explored for use in phase quantization. Essentially, the production of sound can be described most conveniently as passing an excitation through the vocal system which represents the composite characterstics of the glottal pulse, vocal tract and lip-radiation filters. The phase contribution from the excitation can be modeled well by adding together a voicing-dependent random phase $\epsilon(w)$ and a linear component corresponding to the onset time n_0 of the glottal pulse. When combined with the vocal system phase $\Phi(w)$, the complete sine-wave phase synthesis model for the l-th harmonic becomes

$$\theta_l = -n_0 l w_0 + \epsilon(l w_0) + \Phi(l w_0). \tag{4}$$

As a consequence, the success of this representation heavily depends on the accuracy of phase derivation for modeling the vocal system. The most frequently used model is based on the minimum-phase assumption, under which the system phase can be derived by applying a Hilbert

transform to the cepstral envelope of the sine-wave amplitudes. This assumption has proved to be reasonably effective, though further refinements can be achieved by cascading the minimum-phase system with an all-pass filter [4].

Unlike the current STC, the approach taken here is to use a noncausal all-pole filter to model the vocal system. As mentioned earlier, the vocal system represents the composite characteristics of the glottal pulse, vocal tract and lip-radiation filters. It is convenient to combine the glottal pulse filter and lip-radiation filter and represent them as the negative impulse response of an anticausal two-pole filter with transfer function

$$G(z) = \frac{1}{(1 - g_1 z^{-1})(1 - g_2 z^{-1})} \tag{5}$$

where the poles $\{g_1, g_2\}$ lie outside the unit circle. On the other hand, resonant characteristics of the vocal tract can be modeled by means of a causal all-pole filter. Particularly for phase derivation, it suffices to employ a second order filter with transfer function

$$V(z) = \frac{1}{(1 - v_1 z^{-1} - v_2 z^{-2})}. \tag{6}$$

To model the vocal system, these two filters can be combined into a noncausal all-pole filter with the following phase spectrum

$$\Phi(w) = -tan^{-1}\frac{g_1 sinw}{1 - g_1 cosw} - tan^{-1}\frac{g_2 sinw}{1 - g_2 cosw}$$

$$- tan^{-1}\frac{v_1 sinw + v_2 sin2w}{1 - v_1 sinw - v_2 cos2w} \tag{7}$$

To operate the system at 2400 bps, it may not be possible to encode additional information about the filter parameters. Fortunately, good results have been found by using fixed parameters $(g_1, g_2) = (1.1, 1.1)$ and $(v_1, v_2) = (1.515, -0.752)$, which were empirically determined by estimating the long-term-averaged autocorrelation of the speech signals sampled at 8000 Hz.

5. EXPERIMENTAL RESULTS

The suitability of new parametric models introduced above has been evaluated for use in conjunction with the STC vocoders at 2400 bps. Figure 1 displays the experimental arrangement of the proposed coding system. Using an analysis frame length of 22.5 msec, the total number of bits available per frame is 54, with the breakdown according to parameters as shown in Table I. Fourteen subjective loudness scales were represented by an adaptive quantizer whose levels were adjusted to the maximum absolute value within a frame. A 5-bit representation of this maximum absolute value was transmitted as side information regarding the time-varying nature of speech. Next, the individual loudness scales were normalized and uniformly quantized using different degrees of bit resolution. More specifically, two bits were allocated to the first three loudness scales, and three bits to the others. Notably, the phase information is coded implicitly by adding together an mixed excitation phase and a phase contribution due to the vocal system. While the former could be determined by the voicing probability, the latter had to be estimated from the vocal system either through cepstral modeling or through noncausal all-pole modeling. Towards this end, a preliminary experiment was conducted to examine the accuracy of different phase models over 800 frames of voiced vowels including /a/,/e/,/i/,/o/ and /u/. The reference STC algorithm employed in our analysis is the one presented in [1]. The distortion measure applied here is the mean square error between the original waveform and its modeled fit. According to Table II, the noncausal all-pole model outperformed its minimum-phase counterpart with either all-pass compensation included [4] or not [1]. The inadequacy of the minimum-phase assumption appears to result from glottal pulses tending to have rather slowly rising edges but which are terminated by much sharper trailing edges.

Table III presents the comparative performance results for 2400 bps coding of speech in conjunction with different coder structures. For further discussion, the STC described herein and the reference STC [1] are referred to as STC-B and STC-C, respectively. The speech database for these studies consisted of four sentential utterances spoken by two males and two females, each 3 seconds in duration and sampled at 8000 Hz. The performance was evaluated in terms of mel-cepstral distance (MCD) [6] and Bark spectral distance rating (BSDR) [7]. Their results demonstrated that the performance correlated more closely with the results of human preference tests than those obtained by other conventional objective measures. From above results, we can infer that the Bark spectral model is preferred to the cepstral model for its use in amplitude representation, because the former can more closely incorporate the perceptual properties of human hearing. For purposes of comparison, we also included the performance of the speech coder in conforming to the well-established LPC-10e standard [8]. As the table shows, the proposed STC coder yielded substantial improvement over the LPC-10e coder for all test samples. Informal listening tests also indicate that the combined use of a Bark-adaptive amplitude model and a noncausal all-pole phase model allows the STC to deliver synthetic speech of good quality at 2400 bps.

6. CONCLUSIONS

This paper presents some refinements that allow the STC-based speech coder to deliver good quality at 2400 bps. Experimental results demonstrate that Bark spectrum provides an ideal framework for incorporating known features of human hearing in the design of amplitude quantization. In comparison to cepstral-based systems, the Bark-based amplitude coder is preferred because of its ability to achieve a uniform perceptual fit across the spectrum. Algorithms have also been presented that relate the harmonic amplitudes to the Bark spectrum. One enhancement that further increases performance is the use of a noncausal all-pole vocal system that better matches the maximum-phase nature of differentiated glottal pulses.

7. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Low rate speech

coding based on a sinusoidal model," *Advances in Speech Signal Processing*, Chapter 6, S. Furui and M. M. Sondhi, Eds., New York: Marcel Dekker, 1992.

[2] 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.

[3] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Englewood Cliffs, NJ: Prentice Hall, 1989.

[4] X. Sun, F. Plante, B. M. Cheetham, and K. W. Wong, "Phase modeling of speech excitation for low bit-rate sinusoidal transform coding," in *Proc. ICASSP*, pp. 1691-1694, 1997.

[5] W. R. Gardner and B. D. Rao, "Noncausal all-pole modeling of voiced speech," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 1, pp. 1-10, Jan. 1997.

[6] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun., Computation, and Signal Proc.*, pp. 125-128, 1993.

[7] T. Watanabe and S. Hayashi, "An objective measure based on an auditory model for assessing low-rate coded speech," *IEICE Trans. Inf. and Syst.*, vol. E78-D, no. 6, pp. 751-757, June 1995.

[8] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technology*, pp. 40-49, Apr. 1982.

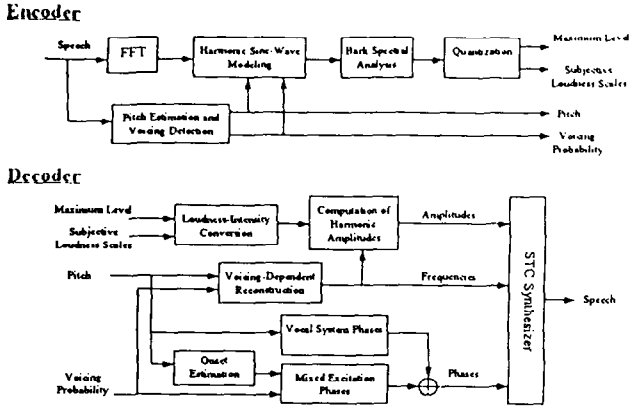

Fig. 1. Block diagram of the Bark-adaptive STC system.

TABLE I
BIT ALLOCATION FOR BARK-ADAPTIVE STC CODERS AT 2400 BPS.

Parameters	Bits
Pitch	7
Voicing Probability	3
Max. Subjective Loudness Level	5
14 Subjective Loudness Scales	39
Total Bits Per Frame	54

TABLE II
THE MEAN SQUARE ERROR (MSE) OF VARIOUS PHASE MODELS.

Phase Models / Vowels	Minimum-Phase	All-Pass Compensation	Noncausal All-Pole
/a/	0.23744	0.19274	0.09560
/e/	0.06708	0.05140	0.04851
/i/	0.34790	0.26249	0.26174
/o/	0.25475	0.20413	0.12009
/u/	0.16993	0.15624	0.10562
Average	0.21542	0.17340	0.12631

TABLE III
MCD/BSDR PERFORMANCES OF VARIOUS SPEECH CODERS AT 2400 BPS.

Coders / Speech	STC-C	STC-B	LPC-10e
Male1	0.311 / 0.095	0.290 / 0.061	0.381 / 0.189
Male2	0.302 / 0.122	0.258 / 0.072	0.386 / 0.159
Female1	0.366 / 0.146	0.329 / 0.075	0.406 / 0.195
Female2	0.338 / 0.128	0.284 / 0.067	0.375 / 0.157