Quality enhancement of sinusoidal transform vocoders

W.-W. Chang
D.-Y. Wang

Indexing terms: Sinusoidal transform coders, Speech synthesis, Bark spectrum

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

1 Introduction

Developments in sinusoidal transform coder (STC) technology have made good quality synthetic speech possible at very low data rates [1–3]. STCs attempt 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 are represented in terms of cepstral coefficients [2] or all-pole modelling [3]. The basic problem with cepstral representation is that the modelling accuracy tends to be uniform across all frequencies and cannot precisely describe the ear's nonlinear responses to frequency selectivity and subjective loudness. Further improvement can only be gained through intelligent exploitation of findings for psychoacoustic studies [4]. This task can be partially achieved by warping the frequency axis to give more prominence to the perceptually more important frequencies [2, 3].

In this paper, we attempt to capitalise more fully on psycho-acoustic knowledge and develop an amplitude coder based on the Bark spectrum [5], instead of based on the properties of the sound production mechanism [2, 3].

As noted elsewhere [2], representing the sine-wave amplitudes by cepstral coefficients has certain advantages towards quantisation. One advantage 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 [6]. Other studies [7, 8], however, indicate the inadequacy of the minimum-phase assumption for modelling voiced speech. This is because glottal pulses tend to have rather slow rising edges which are terminated by much sharper trailing edges. Recognising this, several refinements of the minimum-phase model have been developed that improve the phase accuracy by using either a pulse model (e.g. Rosenberg pulse, LF pulse) or an all-pass filter [7].

Unlike existing STCs, we propose that the vocal system is modelled by a noncausal filter of the all-pole type. The motivation for this representation is twofold. First, it has been shown that noncausal all-pole filters are more appropriate for modelling the vocal system because they take into account the maximum-phase poles of differentiated glottal pulses [8]. Secondly, the minimum-phase assumption is more applicable to versions of STCs which code the sine-wave amplitudes using cepstral coefficients, rather than those using the Bark spectrum. In contrast, the noncausal all-pole approach applies to both representations of sine-wave amplitudes.

2 Harmonic sine-wave model

A promising approach to the parameter quantisation problem lies in the observation that voiced speech, when perfectly periodic, can be represented by harmonic components of its Fourier series decomposition. Other research [1] shows that if the measured frequencies are replaced by integer multiples of pitch frequency, high-quality speech can still be synthesised, provided that the amplitudes and phases are chosen to be harmonic samples of the magnitude and phase spectra, respectively. Viewed from this perspective, 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, and A_l and θ_l are the amplitude and phase of the lth sinusoidal component, respectively.

A low-rate representation is achieved by fitting a set of cepstral coefficients to an envelope of the measured sine-wave amplitudes [2]. For the system with the transfer function $H(z)$, the real cepstrum is defined as the

IEE Proceedings online no. 19982456

Paper first received 9th September 1997 and in revised form 6th July 1998

The authors are with the Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

sequence of coefficients in the power series representation of its log magnitude:

$$c_m = \frac{1}{\pi}\int_0^{\pi} \log|H(w)|\cos(mw)dw \quad 0 \le m \le M-1 \tag{2}$$

The main attraction of cepstral representation is 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 [6]:

$$\log|H(w)| = c_0 + 2\sum_{m=1}^{M-1} c_m \cos(mw) \tag{3}$$

$$\Phi(w) = -2\sum_{m=1}^{M-1} c_m \sin(mw) \tag{4}$$

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 system [9]. The first step is to employ a mixed excitation model, in which the probability of voicing is used to control the harmonic spectrum cut-off frequency. Below this cut-off frequency the excitation phases are made linear, whereas above this cut-off freeequency they are made random on $[-\pi, \pi]$. When combined with the vocal system phase $\Phi(w)$, it has been shown [9] that good quality synthetic speech can be obtained without the need to code the phase information.

McAulay and Quatieri [3] have proposed a technique to estimate the voicing probability by using the degree to which the harmonic model fits the original sine-wave data. This approach is based on the observation that the harmonic fit has poorer accuracy in unvoiced speech than in voiced speech.

3 Perceptual coding of sine-wave amplitudes

In the context of sinusoidal transform coding, 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 frequencies. This inconsistency can be alleviated, to some extent, by warping the amplitude envelope along the perceptually based mel scale before computing the cepstral coefficients [2]. Although 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 psycho-acoustic knowledge. To advance with this, we propose to implement an amplitude coder by using the Bark spectrum [5], rather than the vocal tract envelope as in current STCs [2, 3].

Conceptually, the advantage of the Bark spectrum over the cepstrum is that it more closely emulates several known features of human hearing. As outlined in Fig. 1, the calculation of the Bark spectrum involves the Hertz-to-Bark transformation, critical band filtering, equal loudness pre-emphasis and subjective loudness conversion. First, the power spectrum $X(f)$ is warped along its frequency axis f into the Bark frequency b following the function $b = R(f)$. The resulting warped power spectrum $X(b)$ is then convolved with the simulated critical band masking curve $F(b)$, to yield samples of the excitation pattern $D(b)$ at one-Bark intervals. It typically suffices to use 15 spectral samples of $D(b)$ to cover the 4kHz speech bandwidth.

Consider the vectors **D** and **X** representing the excitation pattern and N-point discrete Fourier transform (DFT) of incoming sound, respectively. For ease of notation, the frequency corresponding to the kth DFT coefficient is referred to as $f_k = kf_s/N$, where f_s denotes the sampling rate. The contribution due to the critical band filtering can be summarised in a matrix $\mathbf{C} = [c_{i,k}]$, where the entry $c_{i,k}$ takes the value of $F[R(f_k) - i]$. With these descriptions, the calculation of excitation pattern **D** can then be formulated as applying the matrix **C** on **X**:

$$\mathbf{D} = \mathbf{C} \cdot \mathbf{X} \tag{5}$$

Loudness is another important attribute of auditory perception, in terms of which sounds can be ranked on a scale extending from quiet to loud. The phenomenon that relates audibility to frequency can be described most conveniently with the equal loudness curve [4]; in turn, this provides an ideal framework for perceptual weighting of spectral energy across the critical bands. For use in the telephone band, a bilinear pre-emphasis filter has been proposed [5] to approximate the equal loudness response.

In the final operation of Bark spectral analysis, a subjective loudness conversion is needed to account for the nonlinear relation between the intensity of sound and its perceived loudness. 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.

Although 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 the Bark spectrum in the synthesis part. This task can be aided by taking advantage of the harmonic modelling assumption, where the sine-wave amplitudes are given by the harmonic samples of the spectral envelope.

The strategies for estimating these harmonic amplitudes may be divided into two steps. First, the Bark spectrum is inversely processed to obtain the excitation pattern following the equal loudness de-emphasis. The next problem is the association of the resulting excitation pattern **D** with harmonic amplitudes, which are exclusively embedded in **X**. An intuitive approach is to obtain the harmonic amplitudes by solving the equation $\mathbf{X} = \mathbf{C}^{-1}\cdot\mathbf{D}$. Unfortunately, however, a unique solution does not exist because the dimension of **D** is less than the number of harmonics. As an example, a typical low-

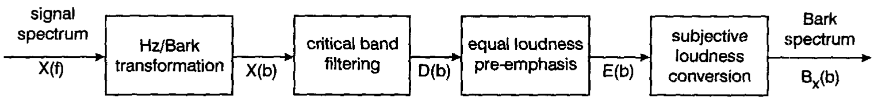

Fig. 1 *Bark spectral analysis*

pitched speaker can have as many as 80 harmonic samples in a 4kHz speech bandwidth, compared to the dimension of 15 in **D**.

Recognising this problem, we propose to develop an analytic method for computing the autoregressive (AR) parameters of the all-pole model, such that its magnitude is a best fit to the underlying excitation pattern. This choice is motivated in part by the success of linear prediction on a warped frequency scale [10], and in part because accurate estimation of AR parameters can easily be found by solving a set of linear equations. We first compute the autocorrelation function by taking an inverse Fourier transform of the excitation pattern. The autocorrelation values are then used to determine the AR parameters of the all-pole model by solving the least-squares Yule-Walker equations [11]. Finally, the sine-wave amplitudes are given by the harmonic samples of the AR modelled fit of $D(b)$.

4 Noncausal all-pole modelling of vocal system

At low rates, more properties of the speech production mechanism need to be explored for use in phase quantisation. Essentially, the production of sound can be described most conveniently as passing an excitation through the vocal system which represents the composite characteristics of the glottal pulse, vocal tract and lip-radiation filters. The phase contribution from the excitation can be modelled well by adding together a voicing-dependent random phase $\varepsilon(w)$ and a linear component corresponding to the onset time n_0 of the glottal pulse [9]. When combined with the vocal system phase $\Phi(w)$, the complete sine-wave phase synthesis model for the lth harmonic becomes

$$\theta_l = -n_0 l w_0 + \epsilon(l w_0) + \Phi(l w_0) \tag{6}$$

As a consequence, the success of this representation heavily depends on the accuracy of phase derivation for modelling the vocal system. The minimum-phase assumption has proved to be reasonably effective, although further refinements can be achieved by cascading the minimum-phase system with an all-pass filter [7].

Unlike in the current STCs, the approach taken here is to use a noncausal all-pole filter to model the vocal system. As mentioned above, 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 [11]

$$G(z) = \frac{1}{(1-g_1 z^{-1})(1-g_2 z^{-1})} \tag{7}$$

where the poles $\{g_1, g_2\}$ lie outside the unit circle.

On the other hand, resonant characteristics of the vocal tract can be modelled by means of a causal all-pole filter. Particularly for phase derivation, it suffices to employ a second-order filter with transfer function

$$V(z) = \frac{1}{(1-v_1 z^{-1} - v_2 z^{-2})} \tag{8}$$

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

$$\Phi(w) = -\tan^{-1}\frac{g_1 \sin w}{1-g_1\cos w} - \tan^{-1}\frac{g_2 \sin w}{1-g_2 \cos w} - \tan^{-1}\frac{v_1 \sin w + v_2 \sin 2w}{1 - v_1 \cos w - v_2 \cos 2w} \tag{9}$$

To operate the system at 2.4kbit/s, it may not be possible to encode additional information about the filter parameters. Fortunately, good results have been obtained by using fixed parameters $(g_1, g_2) = (1.1, 1.1)$ and $(v_1, v_2) = (1.515, -0.752)$; these were empirically determined by estimating the long-term-averaged correlation between coded and natural speech at 8kHz.

5 Experimental results

The suitability of the parametric models introduced above has been evaluated for use in conjunction with the harmonic sine-wave speech model at 2.4kbit/s. Fig. 2 displays the experimental arrangement of the

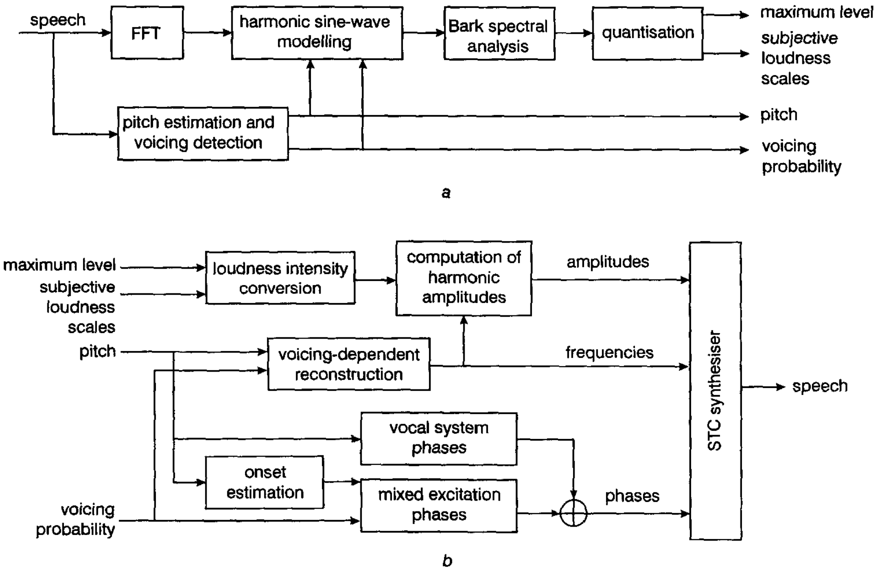

Fig. 2 *Block diagram of Bark-adaptive STC system*
a Encoder
b Decoder

proposed coding system. Using an analysis frame length of 17.5ms, the total number of bits available per frame is 42, with the breakdown according to parameters as shown in Table 1.

Table 1: Bit allocation for Bark-adaptive STC coders at 2400 bit/s

Parameters	Bits
Pitch	7
Voicing probability	3
Maximum subjective loudness level	8
15 subjective loudness scales	24
Total bits per frame	42

15 subjective loudness scales were normalised with respect to the maximum absolute value within a frame. An 8-bit representation of this maximum absolute value was transmitted as side information. We divided the normalised loudness scales into two parts; the first part has the first seven components and the second part has the remaining eight components. Each part is quantised separately using a 4096-level vector quantiser. Notably, the phase information is coded implicitly, by adding together a mixed excitation phase and a phase contribution due to the vocal system. Although the former could be determined by the voicing probability, the latter had to be estimated from the vocal system either through cepstral modelling or through noncausal all-pole modelling.

Table 2: Mean square error of various phase models

Vowels	Minimum phase model	All-pass compensation model	Noncasual all-pole model
/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

Towards this end, a preliminary experiment was conducted to examine the accuracy of different phase models over 800 frames of sustained vowels, including /a/, /e/, /i/, /o/ and /u/. In our analysis, we employed the reference STC algorithm in [2]. The distortion measure applied here is the mean square error between the original waveform and its modelled fit. According to Table 2, the noncausal all-pole model outperformed its minimum-phase counterpart with either all-pass compensation included [7] or not [9]. To elaborate further, some typical waveforms synthesised by the various phase models are shown in Fig. 3. 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.

Computer simulations were then conducted to examine the validity of the Bark spectral model for use in perceptual coding of the sine-wave amplitudes. 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 8kHz. The Bark-based version and the cepstral-based version of the STC are referred to as STC-Bark and STC-cep, respectively. We also compare the proposed system with the well established 4.8kbit/s FS1016 CELP coder [12].

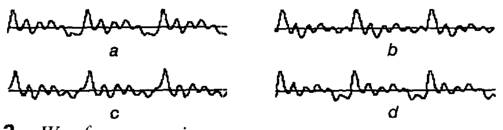

Fig. 3 *Waveform comparisons*
a True phase
b Minimum phase model
c Minimum phase model plus all-pass compensation
d Noncausal all-pole phase model

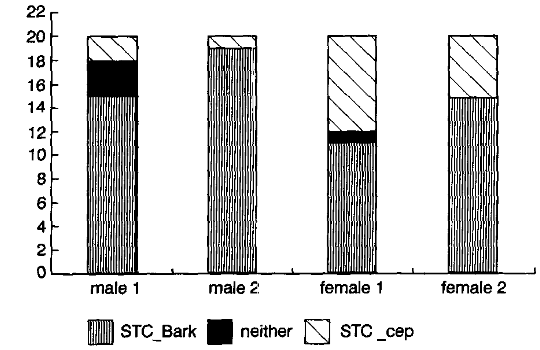

Fig. 4 *A/B test 1: preferences for STC_Bark, STC–cep or neither*

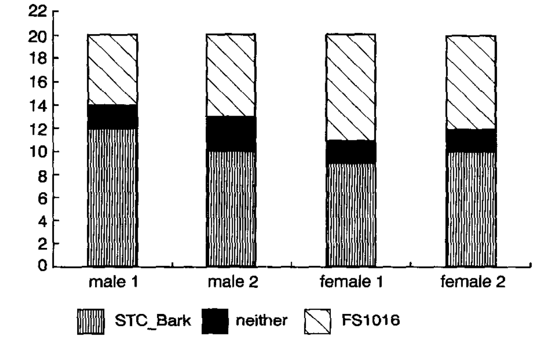

Fig. 5 *A/B test 2: preferences for STC_Bark, FS1016 or neither*

Subjective quality evaluation was accomplished through the use of an A/B comparison test, where 20 participants listened to a number of pairs of speech samples. Each pair has the same content, but one was the STC-Bark version, and the other was either the STC-cep version (in A/B test 1) or the FS1016 version (in A/B test 2). For each pair, the subject selected which one sounded better, or chose neither if they sounded the same. The results of two A/B tests are depicted in Figs. 4 and 5. For each speech sample, the number of subjects who preferred A, B or neither is indicated. The results indicate that the STC-Bark coder is preferred to the STC-cep coder for all test samples, and it performs better or equal to the FS1016 coder at a half rate. Informal listening tests also showed that the combined use of a Bark-adaptive amplitude coder and a noncausal all-pole phase model allows the STC to deliver synthetic speech of good quality at 2.4kbit/s.

6 Conclusions

We have presented refinements that allow the STC to deliver good quality speech at 2.4kbit/s. Experimental results have demonstrated that the Bark spectrum provides an ideal framework for incorporating several known features of human hearing in the design of amplitude quantisation. 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 Acknowledgment

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

8 References

1 McAULAY, R.J., and QUATIERI, T.F.: 'Speech analysis-synthesis based on a sinusoidal representation', *IEEE Trans. Acoust., Speech Sig. Process.*, 1986, **34**, (4), pp. 744–754

2 McAULAY, R.J., and QUATIERI, T.F.: 'Low rate speech code based on a sinusoidal model' *in* FURUI, S., and SONDHI, M.M. (Eds.): 'Advances in speech signal processing' (Marcel Dekker, New York, 1992), Chap. 6

3 McAULAY, R.J., and QUATIERI, T.F.: 'Sinusoidal coding' *in* KLEIJN, W.B., and PALIWAL, K.K. (Eds.): 'Speech coding and synthesis' (Elsevier, Amsterdam, 1995)

4 ZWICKER, E., and FASTL, H.: 'Psychoacoustics' (Springer-Verlag, Berlin, 1990)

5 WANG, S., SEKEY, A., and GERSHO, A.: 'An objective measure for predicting subjective quality of speech coders', *IEEE J. Select. Areas Commun.*, 1995, **10**, (5), pp. 819–829

6 OPPENHEIN, A.V., and SCHAFER, R.W.: 'Discrete-time signal processing' (Prentice–Hall, Englewood Cliffs, New Jersey, 1989)

7 SUN, X., PLANTE, F., CHEETHAM, B.M., and WONG, K.W.: 'Phase modeling of speech excitation for low bit-rate sinusoidal transform coding'. Proceedings of Int. Conf. on ASSP, ICASSP'97, 1997, pp. 1691–1694

8 GARDNER, W.R., and RAO, B.D.: 'Noncausal all-pole modeling of voiced speech', *IEEE Trans. Speech Audio Process.*, 1997, **5**, (1), pp. 1–10

9 McAULAY, R.J., and QUATIERI, T.F.: 'Sine-wave phase coding at low data rates'. Proceedings of Int. Conf. on ASSP, ICASSP'91, 1991, pp. 577–580

10 HERMANSKY, H.: 'Perceptual linear predictive analysis of speech', *J. Acoust. Soc. Am.*, 1990, **87**, (4), pp. 1738–1752

11 RABINER, L.R., and SCHAFER, R.W.: 'Digital processing of speech signals' (Prentice–Hall, Englewood Cliffs, New Jersey, 1978)

12 CAMPBELL, J.P., TREMAIN, T.E., and WELCH, V.C.: 'The federal standard 1016 4800 bps CELP voice coder', *Digital Sig. Process.*, 1989, **1**, (3), pp. 145–155