A HYBRID CODER BASED ON A NEW PHASE MODEL FOR SYNCHRONIZATION BETWEEN HARMONIC AND WAVEFORM CODED SEGMENTS

Nilantha Katugampala and Ahmet Kondoz

Center for Communication Systems Research
University of Surrey, Guildford, GU2 7XH, UK
E-mail: { N.Katugampala, A.Kondoz } @eim.surrey.ac.uk

ABSTRACT

This paper presents a hybrid coder with a new phase model to synchronize harmonic and waveform coded segments, with a target bit rate of 4 kbps. The coder also employs a new technique based on analysis by synthesis to distinguish between stationary and transitional segments. Harmonic excitation is synchronized with the LPC residual by transmitting the location of the pitch pulse closest to the frame boundary and a phase value that represents the shape of the corresponding pitch pulse. The performance of this phase model and the classification technique is evaluated using a hybrid coder. The coder has three modes: scaled white noise excitation colored by LPC for unvoiced, ACELP for transitions, and harmonic excitation for stationary segments. Subjective listening tests show that the coder produces good quality speech and the switching between the modes is transparent.

1. INTRODUCTION

Parametric vocoders based on harmonic and white noise excitation produce highly intelligible speech at bit rates as low as 2.4 kbps [1]. But as the bit rate increases their performance is not asymptotic towards toll quality. This is due to the inadequacy of the harmonic and noise model used by the vocoders, especially at the transitions, e.g. onsets, offsets and plosives. On the other hand waveform coders like ACELP [2] encode the target speech waveform directly and perform better at the transitions. But at low bit rates, waveform coders fail to synthesize stationary segments with adequate quality, because they try to encode even the perceptually unimportant phase information.

Many authors have suggested a hybrid approach to overcome the limitations of a single model, with variations in speech classification, coding methods used and synchronization techniques [3], [4], [5]. Both [3] and [4] use a similar method to ensure signal continuity, where the linear phase deviation between the harmonically synthesized and original speech is measured and the original speech buffer is displaced, such that the waveform coder begins exactly where the harmonic coder has ended. This method needs resetting of the accumulated displacement during unvoiced or silent segments, and may fail to meet the specifications of a system with strict delay requirements.

In [5] signal continuity is preserved by transmitting "alignment phase" for MELP [6] encoded frames, and use of "zero phase equalization" for transitional frames. Zero phase equalization may reduce the benefits of the use of waveform coding by modifying the phase spectrum, and it has been reported that the phase spectrum is perceptually important [7]. Furthermore, zero phase equalization relies on accurate pitch pulse position detection at the transitions, which can be difficult.

In this paper a new phase model for the harmonic coder, Synchronized Waveform matched Phase Model (SWPM) is presented, which preserves both time synchrony and waveform shape between the original and synthesized speech. SWPM does not alter the perceptual quality of the harmonically synthesized speech, and allows ACELP mode to target the original speech waveform without changing the frame boundaries.

2. HARMONIC EXCITATION WITH SWPM

SWPM maintains time synchrony between the original and the synthesized speech by transmitting the Pitch Pulse Location (PPL) closest to each synthesis frame boundary. SWPM also preserves sufficient waveform similarity, such that the switching between the coding modes is imperceptible, by transmitting a phase value, which indicates the Pitch Pulse Shape (PPS) of the corresponding pitch pulse. SWPM needs to detect only the pitch pulses in the stationary voiced segments, which is somewhat easier than detecting the pitch pulses in the transitions as in [5].

SWPM has the disadvantage of transmitting two extra parameters, PPL and PPS, but the bottleneck of the bit allocation of hybrid coders is usually in the waveform coding mode. Furthermore, in stationary voiced segments the location of the pitch pulses can be predicted with high accuracy, and only an error needs to be transmitted. The same argument applies to the shape of the pitch pulses.

2.1. Estimation of Location and Shape of the Pitch Pulses

First, all of the possible pitch pulse locations are determined by considering the localized energy of the LPC residual. Then the locations, which form a possible pitch contour are selected recursively, and a pitch pulse grid is constructed using the pitch period. Finally, the candidate integer pulse position (n_0) closest to the synthesis frame boundary is selected.

Figure 1 depicts a complete pitch cycle of the LPC residual, which includes a selected pitch pulse and the positive half of the wrapped phase spectrum obtained from its DFT. The pulse location, n_0, is taken as the time origin of the DFT, and the phase spectrum indicates that most of the harmonic phases are close to an average value. This average phase value varies with the shape of the pitch pulse, hence we call it Pitch Pulse Shape (PPS). In the absence of a strong pitch pulse the phase spectrum becomes random. The proposed method employs an analysis by synthesis technique in the time domain to estimate PPS, targeting 9 samples

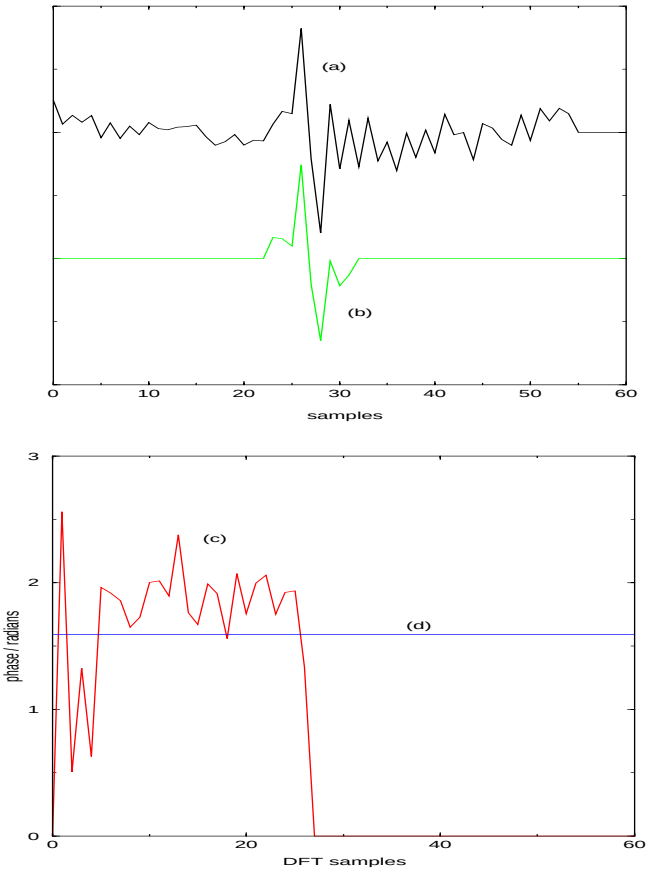

Fig. 1. An Illustration of PPS (a) A complete pitch cycle of the LPC residual, (b) Pitch pulse synthesized using PPS to represent the pulse in (a), (c) Positive half of the phase spectrum obtained from the DFT of the pitch cycle in (a), (d) Estimated PPS

in the vicinity of the detected pitch pulse. A synthetic pulse is generated in an eight times up-sampled domain, i.e. 64 kHz as follows:

$$p_u(n_u) = \sum_{k=1}^{K} a_k \cos(k\omega_u n_u + \alpha_i) \tag{1}$$

where $-40 \leq n_u < 40, \omega_u = 2\pi/8T$, T is the pitch period, K is the number of harmonics, a_k are the harmonic amplitudes, and α_i is the expected PPS given by,

$$\alpha_i = 2\pi i/32 \tag{2}$$

where $0 \leq i < 32$.

Then (3) is used to compute the normalized cross-correlation for all i and j, and the indexes corresponding to the maximum $R_{i,j}$ are chosen as the estimated PPS and the fractional pulse position respectively. Fractional pulse position is important if the pitch pulse is close or at the synthesis frame boundary.

$$R_{i,j} = \frac{\sum_{n=-4}^{4} r(n_0+n)p_j(n)}{\sqrt{\sum_{n=-4}^{4} p_j(n)p_j(n) \sum_{n=-4}^{4} r(n_0+n)r(n_0+n)}} \tag{3}$$

where

$$p_j(n) = p_u(j+8n) \tag{4}$$

where $-8 \leq j < 8, -4 \leq n \leq 4$, $r(n)$ is the LPC residual, and n_0 is the integer pitch pulse location. Figure 1(b) depicts the selected synthesized pulse in the analysis by synthesis process for the pulse shown in Figure 1(a), and Figure 1(d), the straight line across the phase spectrum shows the selected PPS.

2.2. Synthesis using the Generalized Cubic Phase Interpolation

In the synthesis, harmonic amplitudes are interpolated linearly and phases are interpolated cubically, i.e. quadratic interpolation of the frequencies [8]. In [8] phases are interpolated for the frequencies and phases available at the frame boundaries. But in our case the frequencies are available at the frame boundaries and the phases at the pitch pulse locations. Therefore, we use a generalized cubic phase interpolation formula, to incorporate PPL and PPS as follows:

$$\theta(n) = \theta_k + \omega_k n + \alpha n^2 + \beta n^3 \tag{5}$$

where $0 \leq n < N, N$ is samples per frame, θ_k and ω_k are the phase and frequency at the beginning of the frame k respectively, and α and β are given by,

$$\begin{pmatrix} t_0^2 & t_0^3 \\ 2N & 3N^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \theta_{t_0} - \theta_k - \omega_k t_0 + 2\pi M \\ \omega_{k+1} - \omega_k \end{pmatrix} \tag{6}$$

where t_0 is fractional pitch pulse location (PPL), θ_{t_0} is PPS estimated at t_0, and M is the nearest integer to x, where x is given by,

$$x = \frac{1}{2\pi}\left(\theta_k - \theta_{t_0} + \omega_k t_0 + \frac{(\omega_{k+1} - \omega_k)\, t_0^2}{2N}\right) \tag{7}$$

The initial phase θ_k for the next frame is $\theta(N)$, and the above computations should be repeated for each harmonic.

3. ENCODER OVERVIEW

A block diagram of the encoder is presented in Figure 2. The encoder transmits excitation parameters for one of the three modes: harmonic, ACELP, or white noise excitation. LPC parameters are common for all the modes, are estimated every 20 ms, and interpolated every 5 ms in the synthesis process. An initial classification is made based on the tracked energy, low band to high band energy ratio and zero crossing rate, which decides to use either the noise excitation or one of the other modes. The secondary classification based on analysis by synthesis decides to use either the harmonic excitation or ACELP.

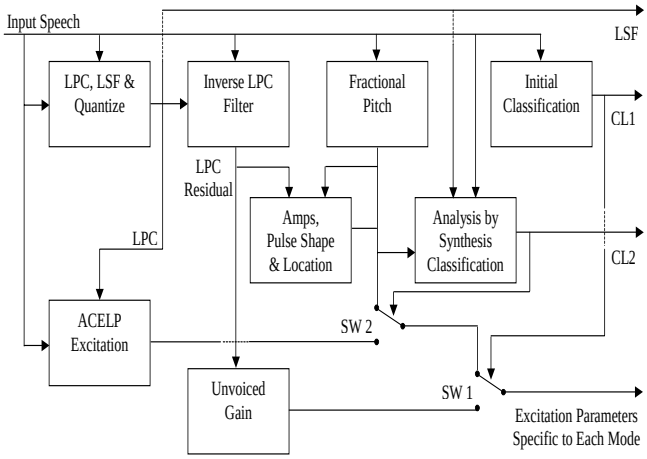

Fig. 2. Block Diagram of the Hybrid Encoder

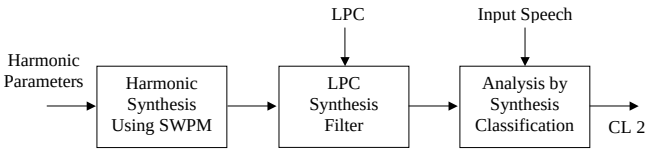

Fig. 3. Analysis by Synthesis Classification

Pitch and harmonic amplitudes are estimated every 10 ms, PPL and PPS are estimated every 20 ms. In the synthesis, amplitudes are linearly interpolated and pitch is updated only for every 20 ms. In ACELP mode, for each sub frame of 10 ms, the adaptive codebook gain and delay, sparse pulse locations and signs of two pulses, and a common pulse gain are transmitted.

In the white noise excited mode the gain estimated from the full band spectral energy is transmitted for every 20 ms. The complicated waveform structure of the unvoiced segments, such as fricatives has no perceptual importance, and can be represented by scaled white noise colored by LPC [9]. If a particular system allows sufficient delay or variable rate transmission this can lead to a significant overall compression, by employing a hybrid coding approach.

For simplicity, details of LPC and adaptive codebook memory update are excluded in the block diagram. The encoder maintains a LPC synthesis filter synchronized with the decoder, and uses the final memory locations for ACELP and analysis by synthesis classification in the next frame. Adaptive codebook memory is always updated with the previous LPC excitation vector regardless of the mode.

4. ANALYSIS BY SYNTHESIS SPEECH CLASSIFICATION

A block diagram of the analysis by synthesis classification process is shown in Figure 3. Analysis by synthesis classification module synthesizes speech using SWPM. For stationary voiced speech, SNR and cross-correlation of the original and the synthesized speech are high, but at the transitions the harmonic model fails, which results in lower cross-correlation and SNR values. The normalized cross-correlation and SNR are computed in both the

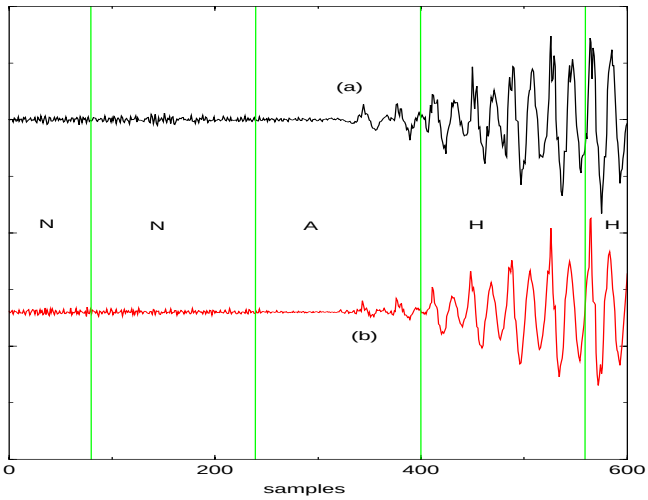

Fig. 4. Synthesized Speech and Classification (a) Original speech, (b) Synthesized speech, N: White noise excitation, A: ACELP, H: Harmonic excitation with SWPM

speech domain and the residual domain for each of the selected pitch cycles in the synthesis frame. The pitch cycles are selected such that they cover the complete synthesis frame. The mode decision between harmonic and ACELP modes is based on the estimated cross-correlation and SNR values, and is biased towards the harmonic excitation such that any one of the four parameters is sufficient to declare the harmonic mode, provided that the selected parameter declares the harmonic mode for all the pitch cycles.

Harmonically synthesized speech at the decoder and speech synthesized by the analysis by synthesis module are similar to original speech. Figure 4 depicts an original speech sample, synthesized speech at the decoder, and the mode used for each synthesis frame.

5. SUBJECTIVE TEST RESULTS

Table 1 shows the bit allocation for different modes, the figures shown within brackets are estimated values and the corresponding parameters are unquantized. The 2 pulses of ACELP sub frames are chosen each from 32 possible locations, either even or odd, covering only the first 64 locations of a sub frame. The pulse gains of the two sub frames are normalized by a common gain, quantised with 3 bits, and then each pulse gain is quantized with 3 bits.

Two pair wise comparative listening tests were carried out to evaluate the performance of a hybrid coder employing the new techniques. For reference, two standard coders were used: ITU 8 kbps G.729 and ITU G.723.1 at the rate of 5.3 kbps. The speech material for the test consists of 8 sentence pairs, 4 from male and 4 from female talkers, filtered by modified IRS filter, and a pair of headphones was used to conduct the test. Fifteen non-expert listeners were used to assess the quality of the hybrid coder as compared against the standard coders.

When compared against G.723.1, there was an overwhelming preference for the hybrid coder. This was due to its cleanness and consistence. However, when compared against G.729 the results were not so conclusive. In fact G.729 was slightly more preferred. This was due to the overall fullness of G.729 and slightly more

Table 1. Bit allocation for a 20 ms frame

Parameters	Harmonic	ACELP	White Noise
LPC	23	23	23
Pitch	(8)	-	-
PPL	(7)	-	-
PPS	(4)	-	-
Amplitudes	$14+14$	-	-
Gain	$(4+4)$	3	5
LTP Delay	-	$7+7$	-
LTP Gain	-	4+4	-
Pulse Locations	-	$5+5+5+5$	-
Pulse Signs	-	$2+2$	-
Pulse Gain	-	$3+3$	-
Mode	2	2	2
Total	80	80	30

metallic character of the hybrid coder. We feel that upon completion of the new developments currently in progress our hybrid coder will produce at least as good quality as G.729 at around half the rate. These improvements include further refinement of transitional sections as offsets, where LPC may be very resonant, and others where more random looking excitation is needed, and optimization of ACELP to match these sections better. Performance of SWPM will also be improved, employing analysis by synthesis techniques in the speech domain, for the voiced speech segments, where the residual pulses become less dominant, e.g. when the LPC spectrum has a very strong formant.

6. CONCLUSIONS

This paper has presented a hybrid coder, which employs synchronized waveform matched phase model (SWPM) in the harmonic mode to preserve signal continuity. A new analysis by synthesis speech classification method to distinguish stationary and transitional segments based on SWPM is also presented. The coder operates in three modes: harmonic, ACELP and white noise excitation. Subjective test results have shown that the hybrid coder designed produces good quality speech, better than 5.3 kbps G.723.1. SWPM shows promising results that would achieve performance similar to G.729, upon completion of the improvements currently in progress.

7. REFERENCES

[1] W. Kleijn and K. Paliwal, *Speech Coding and Synthesis*, Elsevier Science Publishers, Amsterdam, 1995.

[2] C. Laflamme, J. P. Adoul, H. Y. Su, and S. Morissette, "On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, 1990, pp. 177–180.

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

[4] E. Shlomot, V. Cuperman, and A. Gersho, "Combined harmonic and waveform coding of speech at low bit rates," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, 1998.

[5] J. Stachurski and A. McCree, "Combining parametric and waveform-matching coders for low bit-rate speech coding," in *X European Signal Processing Conf.*, 2000.

[6] A. V. McCree and T. P. Barnwell, "Mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 242–250, July. 1995.

[7] Doh-Suk Kim, "Perceptual phase redundancy in speech," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, 2000.

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

[9] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," in *IEEE Workshop on Speech Coding for Telecom.*, 1993, pp. 35–36.