EFFICIENT MIXED EXCITATION MODELS IN LPC BASED PROTOTYPE INTERPOLATION SPEECH CODERS

Charalampos Papanastasiou and Prof. Costas S. Xydeas

Speech Processing Research Laboratory, Electrical Engineering Division,
Manchester School of Engineering, University of Manchester,
Dover St., Manchester M13 9PL, U.K.
Tel: +44 161 2754511, Fax: +44 161 275 4528.
e-mail: C.Xydeas@man.ac.uk

ABSTRACT

This paper presents a new and efficient method for modelling voiced, mixed excitation spectra in Sinusoidal (SC) and Prototype Interpolation Coding (PIC) systems. Speech harmonics are classified as "weak-voiced" or "strong-voiced" by simply examining the short-term residual magnitude spectrum. This information is encoded effectively in terms of fixed width frequency bands and is used to control sets of periodic and random sine wave oscillators which model the short-term mixed excitation nature of speech. In this way the model allows the mixing of periodic and random signal energy on a harmonic basis. The proposed methodology has been used in a 2.4Kbits/sec speech coder, whose recovered speech quality is better than that of the 4.8Kbits/sec DoD standard.

1. INTRODUCTION

In recent years, Sinusoidal [1] (SC) and Prototype Interpolation Coding [2,3] (PIC) systems have been proposed, which provide "communication" quality speech at bit rates below 4Kbits/sec. In principle, PIC and SC schemes are related through the general signal synthesis equation:

$$x(i) = \sum_{k=0}^{K} A_k(i) \cos\left(\vartheta_k(i) + \phi_k\right) \qquad (1)$$

which can be used to represent: *i)* the speech signal s(i) itself (i.e. x(i)=s(i)), or *ii)* the excitation signal e(i) (i.e. x(i)=e(i)) in a "source-filter" speech synthesis process. In Equation 1, i is the sampling instant and, $A_k(i)$ and $\Theta_k(i) = \vartheta_k(i) + \phi_k$ represent the instantaneous amplitude and phase values respectively, of the kth cosine term $\cos\left(\Theta_k(i)\right)$.

The above systems operate on a frame by frame basis and the $A_k(i)$, $\Theta_k(i)$ functions are determined using interpolation techniques on the A_k amplitude and ω_k frequency values of each coding frame. The A_k amplitudes can be defined: *i)* from the "peaks" of the speech short-term magnitude spectrum, as in SC, or *ii)* as the magnitude spectrum of a pitch segment, which is the case in PIC systems. The instantaneous amplitude information $A_k(i)$ is usually defined using linear interpolation. Furthermore, the ω_k cosine frequencies are assumed to be "harmonically" related and to evolve linearly with time. The $\Theta_k(i)$ instantaneous phase function is then defined as the integral of $\omega_k(i)$. Notice that the exact phase information of the signal is not preserved in coders operating in the region of 2.4Kbits/sec.

In the general synthesis model of Equation 1, three significant factors determine the quality of the recovered speech signal: *i)* the parametric representation and quantisation of the A_k amplitude information [4], *ii)* the rate at which model parameters are updated in Equation 1, and *iii)* the "mixed-excitation" mechanism, which introduces, when necessary, the appropriate amount of randomness in the otherwise "periodic" speech reconstruction process. Several random mixed excitation procedures that operate within the synthesis framework of Equation 1 have been reported in the literature. These involve: *a)* extending Equation 1 in order to provide appropriate random signal components at certain spectral areas [1],[3],[5], and *b)* a multiband approach [6], where the signal spectrum is represented as the combination of non-overlapping periodic and random spectral bands.

The use of mixed excitation in representing voiced speech, removes the reverberation and tonality artifacts which are associated with the "harmonic" signal synthesis manifestation of Equation 1. Notice that these mixed excitation mechanisms can be also used to model the random nature of unvoiced speech and perform as well as conventional unvoiced excitation mechanisms.

This paper presents a novel and perceptually powerful method for modelling voiced mixed excitation spectra. This method is applied to an LPC-based PIC synthesis model, within the context of a high speech quality 2.4Kbits/sec codec which has been developed in Manchester (MAN-LPC-PIC) [7]. The codec operates on 20msecs duration input speech frames and each frame is then classified as voiced or unvoiced. PIC principles are used to synthesise voiced speech, while employing the proposed mixed excitation technique. Thus, the system classifies signal harmonics, on a short term basis, as "strong-voiced" or "weak-voiced". This information is represented efficiently in terms of fixed width frequency bands and is used at the receiver to control periodic and random oscillators whose frequencies and amplitudes are defined according to the mixed excitation model parameters. Unvoiced speech is recovered as the output of an LPC filter

whose input is white gaussian noise that is frequently scaled by the RMS value of the original LPC residual signal.

Section 2 of this paper discusses the efficient classification of signal harmonics as "strong" or "weak" voiced and the representation of this information. The new hybrid excitation LPC-PIC speech synthesis process is presented in Section 3, whereas its performance when used in the MAN-LPC-PIC codec, is presented in section 4. Finally, section 5 of this paper contains concluding remarks.

2. CLASSIFICATION OF HARMONICS

Consider that the nth, 20msecs input speech frame has been classified as voiced (V_n=1) and that the pitch period value attached to this frame is P_n. Classification of the ω_k^n k=1,2,...,P_n/2 harmonics into "strong" or "weak" voiced is performed by examining the magnitude spectrum of the 20msecs residual signal derived via inverse LPC filtering the nth input frame. Conventional harmonic/band classification methods used in SC rely on accurate, "non-integer" pitch estimates. This allows the accurate modelling of the "periodic components" in the signal's magnitude spectra, and thus the formulation of an error/SNR function which reflects the difference between the "actual" and the "periodic-model" spectra. Classification of spectral bands into "voiced" or "unvoiced" is then performed using these error/SNR functions.

In contrast, MAN-LPC-PIC employs integer pitch estimate values and defines $\omega_k^n = k(2\pi / P_n)$, k=1,2,...,$P_n$/2. The short-term residual magnitude spectrum is clipped and a set of dominant peaks are obtained using a "peak-picking" process. Harmonics ω_k^n are then "associated" to dominant peaks, see Figure 1. A peak "associated" with a harmonic ω_k^n enables the classification of that harmonic as "strong" voiced (hv_k^n=1). The absence of a dominant peak leads to a "weak" voiced classification (hv_k^n=0) for that harmonic.

Furthermore, in order to reduce the bit rate allocated to hv_k^n, the bandwidth of the input signal is divided into M fixed size bands BD_j and a strong/weak voiced flag Bhv_j is assigned to each band. Bands are classified as "strong" (Bhv_j^n=1) or "weak" (Bhv_j^n=0) voiced, using a majority decision rule on the hv_k^n classification values of the harmonics ω_k^n contained within each frequency band. Further restrictions can be imposed on the strong/weak voiced classification of harmonics. For example, the first L bands are always "strong" voiced, i.e. Bhv_j^n=1 for BD_j with j=1,2,...,L, whereas the last H bands are always "weak" voiced, i.e. Bhv_j^n=0 for BD_j with j=H, H+1,...,M. The remaining spectral bands can be "strong" or "weak" voiced. Following this scenario, subjective listening tests indicated that the hv_k^n information is efficiently represented by a scheme which employs 500Hz bands, and L=1, H=7 when V_{n-1}=0, or L=2, H=8 in the case of V_{n-1}=1, (i.e. 5 bits are allocated to quantise the hv_k^n information).

Alternatively, the hv_k^n information can be represented economically by an adaptive dual band approach. In this case, three bands defined by four candidate cut-off frequencies (i.e. {680, 1320, 2040, 2720Hz} if V_{n-1}=0 or {1320, 2040, 2720, 3400Hz} when V_{n-1}=1) are examined sequentially starting with the lower band. Using again a majority rule on hv_k^n, bands are classified and when a "weak" voiced band is found, the lower boundary of this band determines a cut-off frequency ω_{Fc} and the process then stops. At the decoder, harmonics ω_k^n<ω_{Fc} are considered as "strong" voiced whereas those which are larger than ω_{Fc} are treated as "weak" voiced.

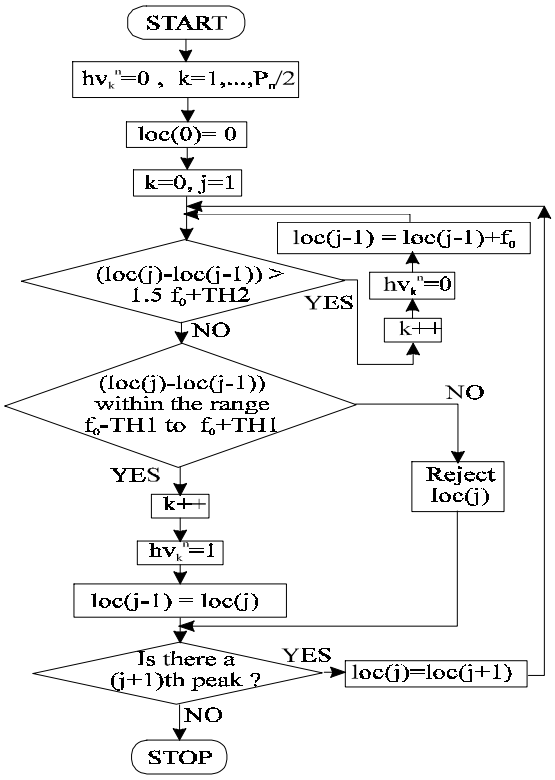

Figure 1. Classification of each harmonic as voiced (hv_k^n=1), or "mixed voiced" (hv_k^n=0). Notice that harmonic indices k do not always correspond to the magnitude spectrum peak indices j. loc(j) is the location of the jth dominant peak, $f_0 = (1/P_n) f_s$, f_s is the sampling frequency, $TH1 = 0.15 f_0$ and $TH2 = (1.5/P_n) f_0$. (the symbol ++ denotes increase by one)

3. MIXED EXCITATION SYNTHESIS PROCESS

A diagram of the voiced excitation synthesis process is shown in Figure 2, where the term "hybrid harmonic oscillator" is used to highlight the mixed (periodic/random) nature of the oscillators. When $hv_k^n = 1$, the contribution of the kth harmonic to the synthesis process is $\hat{A}_k^n \cos\left(\int \omega_k^n(i)di\right)$, with $\hat{A}_k^n$ being the decoded amplitude value of the ω_k^n frequency. On the other hand, when $hv_k^n = 0$, the frequency of the kth harmonic is slightly dithered, its amplitude $\hat{A}_k^n$ is reduced to

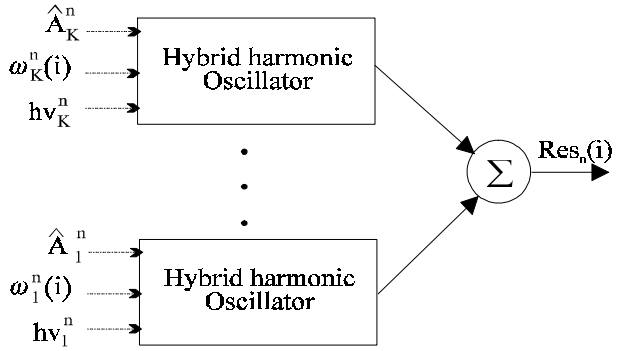

Figure 2. Schematic diagram of the voiced excitation synthesis process.

$\left(\hat{A}_k^n / \sqrt{2}\right)$, and random cosine terms are added symmetrically alongside the kth harmonic ω_k^n (see Figure 3). The number NRS of these random terms is defined as the number of 50Hz intervals that can fit within the fundamental frequency band (i.e. f_s/P_n). The amplitudes of the NRS random components are set to $\left(\hat{A}_k^n / \sqrt{2\,NRS}\right)$. Thus, when hv_k^n=0, half of the signal energy initially associated with the kth harmonic is now allocated to a random signal component, whereas the remaining signal energy is represented by a periodic component. The phases of the "random" oscillators are selected randomly from the [-π, +π] region, at pitch period intervals.

In the transitions from unvoiced to voiced frames, the initial phase for each harmonic is set to zero. Phase continuity, however, is preserved across the boundaries of successive voiced PIC interpolation intervals.

Notice that the excitation synthesis process (voiced or unvoiced) is performed twice in the MAN-LPC-PIC system. Once using the parameters derived for the current nth frame and then using the model parameters of the previous (n-1)th frame. However, when decoding voiced frames, the same $\omega_k^n(i)$ function is employed in both synthesis procedures. The resulting residual signals, $Res_n(i)$ and $Res_{n-1}(i)$, are used as inputs to the corresponding LPC synthesis filters defined for the nth and (n-1)th input frames. The two LPC synthesised speech waveforms are then weighted by the Hamming window functions $W_n(i)$ and $W_{n-1}(i)$ and they are added, to yield the recovered speech. Thus, the overall synthesis process for successive voiced frames can be described by:

$$S_n(i) = W_n(i)\sum_{\lambda=1}^{\Lambda} H^n\left(\omega_\lambda^n(i)\right)\tilde{A}_\lambda^n \cos\left[\Theta_\lambda^n(i) + \varphi^n\left(\omega_\lambda^n(i)\right)\right] + W_{n-1}(i)\sum_{\lambda=1}^{\Lambda} H^{n-1}\left(\omega_\lambda^n(i)\right)\tilde{A}_\lambda^{n-1} \cos\left[\Theta_\lambda^n(i) + \varphi^{n-1}\left(\omega_\lambda^n(i)\right)\right] \quad (2)$$

where Λ is the total number of "periodic" and "random" oscillators and $\tilde{A}_\lambda^n$ represent their amplitude values. $H^n\left(\omega_\lambda^n(i)\right)$ is the frequency response of the nth frame LPC

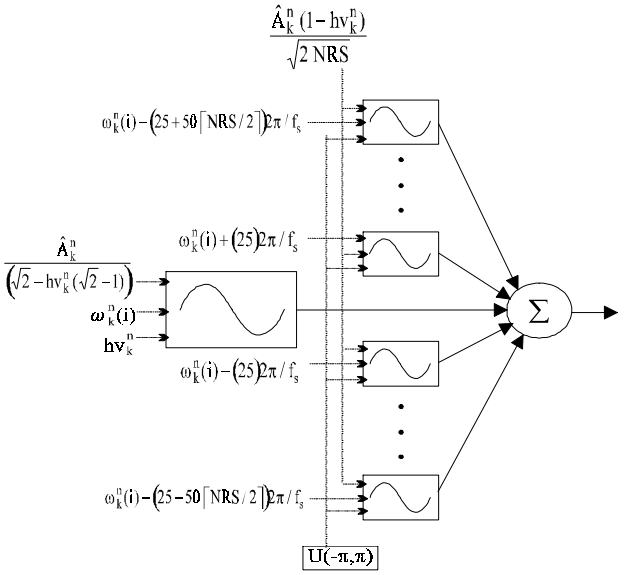

Figure 3. Schematic diagram of the kth "hybrid harmonic oscillator".

synthesis filter calculated at the $\omega_\lambda^n(i)$ instantaneous harmonic frequency and $\varphi^n\left(\omega_\lambda^n(i)\right)$ is the associated phase response of this filter. Notice that the frequency and phase functions $\omega_\lambda^n(i)$ and $\Theta_\lambda^n(i)$ are defined for each sampling instants i, where i ranges from the middle of the nth frame to the middle of the (n-1)th frame.

The above speech synthesis process introduces two "phase dispersion" terms, i.e. $\varphi^n\left(\omega_\lambda^n(i)\right)$ and $\varphi^{n-1}\left(\omega_\lambda^n(i)\right)$, which effectively reduce further the degree of pitch periodicity in the recovered signal. In addition, this "double synthesis" arrangement, followed by a weighted overlap-add process, ensures an effective smooth evolution on a sample by sample basis, for the LPC speech spectral envelope.

Notice that in Equation 2, the amount of overlap is controlled by the $W_n(i)$ and $W_{n-1}(i)$ functions. In the case of adjacent frames of the same voicing status, the overlap region extends from the middle of the (n-1)th frame to the middle of the nth frame, whereas in voicing transitions, the overlap region is reduced to 10 msecs. In the latter case, abrupt unvoiced to voiced transitions are modelled with a smooth passage from "random" to "periodic" speech and the exact voicing onset point is lost. However, informal listening tests have shown then this approximation is perceptually acceptable.

The final stage of the MAN-LPC-PIC speech synthesis process includes a postfiltering operation, which enhances output speech quality.

4. PERFORMANCE ASSESSMENT

The performance of the proposed mixed excitation synthesis process has been assessed in terms of informal listening tests. These tests have clearly demonstrated the value of the mixed

excitation model (hv_k^n=1 or 0) over the "binary" excitation model where hv_k^n=1.

Figure 4, which provides an example of short-term LPC envelope and residual signal magnitude spectra, highlights this point and illustrates the superiority of the mixed excitation model in the frequency region $3\pi/4$ to π.

It is also important to notice in this example, the influence of the random excitation components, which the system employs in the $[3\pi/4, \pi]$ region, on the "strong" voiced harmonics $\omega_k^n < 3\pi/4$. Recall that the frequencies of the "random" oscillators are located relatively near to ω_k^n harmonics classified as hv_k^n=0, and their phases are randomised at pitch intervals. This effectively adds a pitch related random component in the synthesis process. Figure 5 depicts the long-term magnitude spectrum of the output signal obtained from a "random" oscillator of frequency CF=2.5KHz, whose phase is randomised every 40 samples (i.e. F_1=200 Hz). As expected, the spectrum exhibits random-like characteristics in the frequency band [CF-F_1/2, CF+F_1/2]. Furthermore, minima are observed at F_1Hz intervals spaced symmetrically around CF. When such a "random" oscillator is involved in the synthesis Equation 2, its angular frequency $\omega_r^n = (CF / f_s)2\pi$ is located near to a "weak" (i.e. hv_k^n=0) ω_k^n, harmonic and the spacing $(F_1 / f_s)2\pi$ between the above spectral minima is equal to ω_0^n. Consequently, harmonic spectral peaks in the entire magnitude spectrum are affected by the spectral minima of the random oscillators. As a result, the spectral peaks of strong voiced harmonics are not separated as clearly as in the binary excitation case, see Figure 4, (C) and (D).

5. CONCLUDING REMARKS

This paper presented an efficient mixed excitation LPC-PIC model. Classification of harmonics as "strong" or "weak" voiced is achieved, without using accurate, non-integer pitch estimate values, by a simple process that examines the "distance" between adjacent dominant peaks in the short-term residual magnitude spectrum. This information is accurately encoded with 2 or 5 bits, in terms of fixed width frequency bands and is used in a frequency selective mixed excitation synthesis process which: *a)* operates on a harmonic basis allowing for both periodic and random components to exist at specify spectral areas, and *b)* provides a "fuzzy" time evolution of "weak" voiced harmonic frequencies. The proposed mixed excitation methodology led to 2.4Kbits/sec MAN-LPC-PIC codecs which provide subjectively better speech quality, than the 4.8Kbits/sec 1016 DoD standard [8].

References

[1] McAulay and T.F. Quatieri, *'Speech Analysis/Synthesis Based on a Sinusoidal Representation*", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-34, No. 4, pp. 744-754, August 1986.

[2] W. B. Kleijn, "*Encoding Speech Using Prototype Waveforms*", IEEE Trans. Speech Audio Processing, Vol. 1, pp. 386-399, 1993.

[3] W. B. Kleijn and J. Haagen, *"A Speech Coder Based on Decomposition of Characteristic Waveforms"*, IEEE Proc. ICASSP-95, pp. 508-511, Detroit, May 1995.

[4] C. Papanastasiou and C.S. Xydeas, "*Advances in Prototype Interpolation Coding at 1.5 and 2.4Kb/sec.*", Proc. of IEEE Inter. Workshop on Intelligent Signal Proc. and Comm. Systems, pp. 731-735 Singapore, November 1996.

[5] G. Yang, G. Zanellato and H. Leich, *'Band Widened Harmonic Vocoder at 2 to 4 kbps*", IEEE Proc. ICASSP-95, pp. 504-507, Detroit, May 1995.

[6] D. W. Griffin and J. S. Lim, *'Multiband Excitation Vocoder*", IEEE Transactions on Acoustics, Speech and Signal processing, vol. 36, pp. 1223-1235, 1988.

[7] C. Papanastasiou, "*LPC-Based Pitch Synchronous Interpolation Speech Coding*", Ph.D. Thesis, Manchester School of Engineering, University of Manchester, 1996.

[8] J.P. Campbell, T.E. Tremain and V. C. Welch, *"The DoD 4.8 KBPS Standard (Proposed Federal Standard 1016)"*, Advances in Speech Coding, Kluwer Academic Publishers, pp. 121-134, 1991.

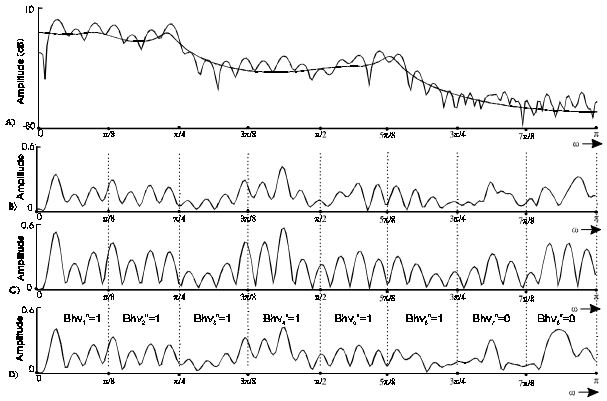

Figure 4. A) Short-term speech magnitude spectrum and LPC spectral envelope ($\log_{10}$ domain), and the short-term magnitude spectra of the corresponding: B) residual segment, C) excitation segment obtained using the binary excitation model, and D) excitation segment obtained using the mixed excitation model.

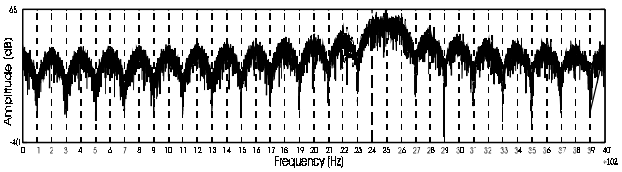

Figure 5. Long-term magnitude spectrum of the output signal of a "random" oscillator at 2.5KHz, whose phase is being randomised every 40 samples (200Hz).