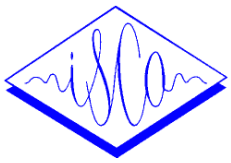

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SINE WAVE EXCITED LINEAR PREDICTIVE CODING OF SPEECH

Suat Yeldener, Ahmet M. Kondoz, Barry G. Evans
Department of Electronic and Electrical Engineering
University of Surrey
Guildford, Surrey, GU2 5XH.
UK.

ABSTRACT

The choice an algorithm of speech coding is very important to achieve high quality speech at low bit rates. Speech can be modeled using LPC and Sinusoidal Transform Coding (STC). In LPC, it leads to CELP type coders [1][2]. In CELP, during vector quantization of the excitation, all components are matched as a single vector. This produces background noise and hence roughness below 4.8 kbits/s. In STC [3], on the other hand, the model parameters (phase and frequency) are very sensitive to quantization errors. This affects the performance of this system under channel errors even though it produces high quality speech at low bit rates. In our previous work, we used sine wave components to represent the CELP excitation [4] and LPC residual waveform [6] which both are capable of synthesizing speech without the artifacts common to model-based speech system. In this paper, we present the sine wave excited linear prediction (SWELP) speech model which has been found to be robust in the presence of quantization noise in speech. These characteristics make the model particularly useful in the development of high quality speech coding system at low bit rates.

1. INTRODUCTION

The most promising coding schemes at bit rates of 9.6 to 4.8 kbits/s are Analysis By Synthesis (ABS) type coders such as CELP [1][2]. CELP [1] has been dominating 9.6 to 4.8 kbits/s region during the past 3 to 4 years. However, below 4.8 kbits/s, due to its large required vector dimensions, it has noticeable quantization noise. For low bit rate speech coding, we are looking for a speech coder which is not very costly, thus we need a scheme which is simple to implement on a single DSP chip. Although, Base Band CELP (CELP-BB) satisfies this requirement [2], below 4.8 kbits/s it produces back ground noise and hence roughness in speech quality. In addition to simplicity of a coder, we are also looking for a speech coder producing high quality speech at low bit rates (2.4 - 4.8 kbits/s).

In our previous work, were presented the CELP excitation [4] and LPC residual [6] by the component of sine waves which, in both cases, produced an improvement in speech quality at low bit rates. In this paper we present results of a coder called Sine Wave Excited LPC (SWELP) which we believe is capable of providing high quality reproduction of both clean and noisy speech without the buzziness and severe noise degradation typically associated with vocoded speech.

2. SWELP SPEECH CODING SYSTEM

One approach to the problem of representation of speech signals is to use the speech production model in which speech is viewed as the result of passing a LPC residual waveform through a linear LPC filter that models the resonant characteristics of the vocal tract. In this speech model, the LPC excitation waveform is assumed to be composed of sinusoidal components of arbitrary amplitudes, frequencies and phases. It is called a Sine Wave Excited Linear Prediction (SWELP) vocoder and operates as follows:

In the analysis, a block of speech, $S(n)$ is first LPC analysed to obtain the LPC coefficients. These coefficients are then quantized. The quantized coefficients are used to form an inverse filter to derive the LPC excitation in sub-blocks $r(n)$. The frequency spectrum of the obtained LPC excitation is then analysed to determine the sine wave components (amplitudes, frequencies and phases), using a 512 point FFT and a hamming window with a minimum width of 2.5 times the average pitch, for accurate peak estimation. However, it is impossible to code all of the

uniformly spaced samples of the spectrum at low bit rates. Therefore, in [3], a method was developed through which the spectrum can be represented by a limited number of peaks. The number of peaks is also a function of the measured fundamental frequency. The locations of the peaks are estimated by simply searching for a change of slope from positive to negative in the uniformly spaced samples of the short-time Fourier transform magnitude. The amplitude and phase components (modula 2π) of the sine waves are given by the appropriate samples of the high resolution FFT corresponding to the chosen frequency locations. The computed LPC coefficients and the set of amplitudes, frequencies and phases are then transmitted.

In the synthesis, the received set of sine wave components is used to generate the sine waves for each tone of speech. The generated sine waves are then added together to form the LPC excitation, $r_m(n)$ as,

$$r_m(n) = \sum_{k=0}^{L_m} A_k^m \cos(n\omega_k^m + \Phi_k^m) \qquad (2.1)$$

where L_m is the number of sine waves and A_k^m, ω_k^m and Φ_k^m are the amplitude, frequency and phase, respectively, for the k^{th} sine wave component in the m^{th} frame. The final quantized speech, $\hat{S}(n)$ is then obtained by passing the recovered LPC excitation through the LPC filter.

Due to the time varying nature of the parameters, it is important that these parameters are interpolated at the frame boundaries. In SWELP [4], therefore, we used the LPC filter to interpolate the sine wave components. This way, all the discontinuities were eliminated from the recovered speech with the cost of coding the LPC coefficients. In this case, 7 or 8 LPC coefficients were found to be sufficient for smoothly interpolating the sine wave components.

The SWELP speech model led to synthetic speech that is essentially perceptually indistinguishable from the original, offering potential for high quality speech coding system. The question arises as to whether the parameters of the sinusoidal model can be coded at low data rates (2.4 - 4.8 kbits/s) to result in a high quality speech compression system. Since the parameters of the SWELP speech model are the LPC coefficients, amplitudes, frequencies and phases of the underlying sine waves, and since for a typical low pitched speaker (50 Hz), there can be as many as 80 sine waves in a 4 kHz speech bandwidth, it is not possible to code all of the parameters directly for low bit rate speech. Therefore, speech specific properties are introduced to reduce the size of the parameter set to be quantized for low bit rate speech.

3. AMPLITUDE ENVELOPE ESTIMATION

In the speech production model, the speech waveform $S(n)$ is assumed to be the output of passing a glottal excitation waveform, $r(n)$ through a linear time varying filter that models the characteristics of the vocal tract. If the time varying impulse response of the LPC filter is $h(n)$, then

$$S(n) = \sum_{m=0}^{n} h(n-m)\, r(m) \qquad (3.1)$$

Previously, it was shown that the LPC excitation be represented in terms of a sum of sine waves as shown in equation (2.1). The question arises now as to whether the magnitude envelope of this excitation spectrum can be represented by a reduced number of parameters. For this purpose, if the LPC filter transfer function is,

$$H(\omega) = G(\omega)\, e^{j\alpha(\omega)} \qquad (3.2)$$

then using (2.1) and (3.2) in (3.1), this results in the speech model,

$$S(n) = \sum_{k=1}^{L_m} M_k \cos(n\omega_k + \psi_k) \qquad (3.3)$$

where M_k is the amplitude and ψ_k is the phase of the speech spectrum. The magnitude envelope, $\hat{M}_k$ can be derived using an all-pole model to fit an envelope to the measured peaks of the speech spectrum. The magnitude, A_k and phase, ϕ_k envelope of the LPC excitation spectrum are then determined as,

$$A_k = \frac{\hat{M}_k}{G(\omega)} \qquad (3.4)$$

and

$$\phi_k = \psi_k - \alpha(\omega) \qquad (3.5)$$

Therefore, the LPC excitation amplitudes are not necessary for transmission except their energy level; they could be obtained using the LPC coefficients which have already been transmitted.

4. FREQUENCY AND PHASE MODELING

The next step to reduce the number of parameters to be quantized is to force the speech sequences to be in harmonics. In this case, the

sine wave frequencies can be represented by a multiple of one fundamental frequency. For this purpose, a pitch extraction algorithm was developed. With this strategy, the pitch period was calculated by maximizing the equation,

$$E(p) = MAX\left\{\sum_{i=0}^{N-1} h(i+p)\, r(i)\right\} \quad (4.1)$$

where p is the pitch lag varying from 20 to 90 samples, $h(i)$ is a buffer consisting of the previous LPC excitation samples and $r(i)$ is the smaller sized LPC excitation records of N samples which contain overlaping samples from the previous block. The fundamental frequency, f_o was then calculated as $f_o = 1/p$. However, this straight forward approach introduces pitch doubling effects [6]. For this purpose, in the peak-picking algorithm, the total energy of the estimated peaks (E_p) was calculated and compared with the total energy of the harmonic peaks (E_h). In this case, a threshold was defined. If the ratio of the computed energies, E_h/E_p remained below the defined threshold, the fundamental frequency was then recomputed as $\hat{f}_o = f_o/2$. If the ratio remained above the threshold, the pitch frequency was kept same as computed. This way, the pitch doubling effect is eliminated. The mean squared error between the original and generated harmonic LPC excitation was computed and shown in figure 1.

The computed harmonic set is perceptually best fitted to the measured sine waves. With this strategy, coding of individual sine wave frequencies is avoided. A new set of sine wave amplitudes and phases is then obtained by sampling an amplitude and phase envelope at the pitch harmonics.

Proper representation of phase in a sinusoidal speech synthesizer is crucial to good speech quality. Unlike the magnitude spectrum, the phase spectrum need only be matched at the harmonics. Previously a predictive phase model was developed. The actual phase of the LPC excitation with frequency ω_k can be calculated as,

$$\phi_k = -n_o \omega_k + \epsilon_k \quad (4.2)$$

where n_o is the location in time of the onset of a pitch pulse and ϵ_k is the phase error compensation component [4][6]. In [5] ϵ_k is transmitted for a few harmonics and assumed to be zero for the rest. In a system with no voicing decision, however, this adds noise to voiced speech and makes unvoiced speech buzzy when phase is predicted. ϵ_k's may be coded by replacing ϵ_k with a random vector, $V_k(i)$, $1 \leq i \leq CB$, selected from a code-book of CB code-words. Code-word selection consists of an exhaustive search to find the code-word yielding the least mean squared error (MSE). The MSE between two sinusoids of identical frequency and amplitude but differing in phase by an angle Δ_k is $A_k[1 - cos(\Delta_k)]$. The code-word is chosen to minimize,

$$E(i) = \sum_{k=1}^{L_m} A_k^2[1 - cos(\epsilon_k - V_k(i))] \quad (4.3)$$

Since phase error compensation components in a given spectrum tend to be uncorrelated and uniformly distributed, the code-words are constructed from uniformly distributed noise sequences.

The resulting parametric SWELP model was incorporated into a non-real time program and found to dramatically improve the quality of the output speech. Although, improvements are most noticeable at low rates where no phase coding is possible, the phase locking technique can be used to perform the high frequency reconstruction in those cases where not all of the base band phases are coded. The waveforms of LPC residual signal (original and reconstructed) are shown in figure 2. In informal listening tests, results suggest that very good quality can be obtained at 4.8 kbits/s and natural sounding speech could be obtained at 2.4 kbits/s. Furthermore, since the complete system depends on the LPC coefficients, measured average fundamental frequency and parametric phase and spectral level of an amplitude envelope, operation at rates from 2.4 to 9.6 kbits/s could be obtained with variation in speech quality.

5. CONCLUSION

A number of new developments in SWELP system which yield high intelligibility and quality at a variety of rates have been described. Depending on detailed bit allocation rules, operation at rates from 9.6 to 2.4 kbits/s could be obtained. At rates above 4.8 kbits/s, enough phase information could be coded so that very good speech could be obtained. In order to preserve the naturalness at rates below 4.8 kbits/s, a synthetic spectral envelop estimator and phase model was described which produce natural sounding speech at rates as low as 2.4 kbits/s.

6. REFERENCES

[1] M. R. Schroeder, B. S. Atal, "Code-excited Linear Prediction (CELP): High quality speech at very low bit rates", Proc. of ICASSP-87, pp.1649 - 1652.

[2] A. M. Kondoz, B. G. Evans "CELP base-band coder for high quality speech coding at 9.6 to 2.4 kbits/s" Proc. of ICASSP-88 pp.159-162.

[3] R. J. McAulay and T. F. Quatieri "Speech analysis/synthesis based on a sinusoidal representation" IEEE trans. ASSP-34, pp.744-754 (August 1986).

[4] S. Yeldener, A. M. Kondoz, B. G. Evans "A variable rate speech compressor for mobile applications", International Mobile Satellite Conference, Ottawa, Ontario, Canada, June 1990, pp.690-695.

[5] R. J. McAulay and T. F. Quatieri, "Phase modeling and its application to STC", ICASSP-86, Tokyo, Japan, p. 1713, April 1986.

[6] S. Yeldener, A. M. Kondoz, B. G. Evans, "Speech coding based on the sinusoidal representation of LPC residual" Int. Symp. on Digital signal processing and its applications, Australia, August 1990.

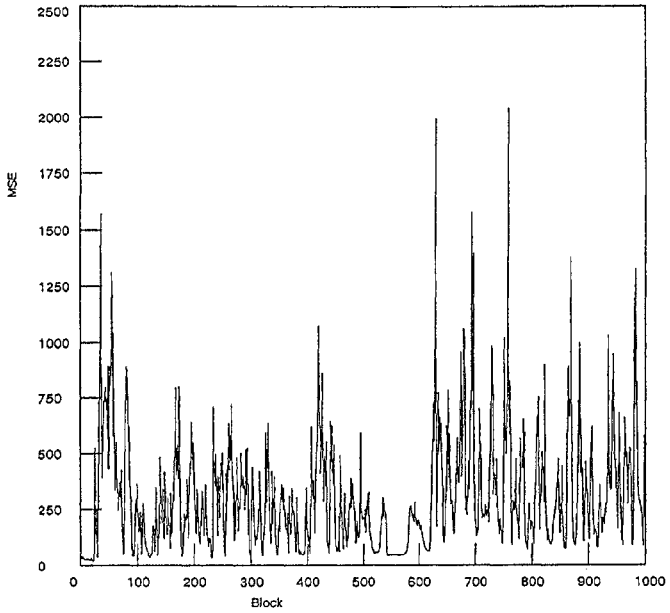

Figure 1 The mean squared error between original and generated LPC excitation.

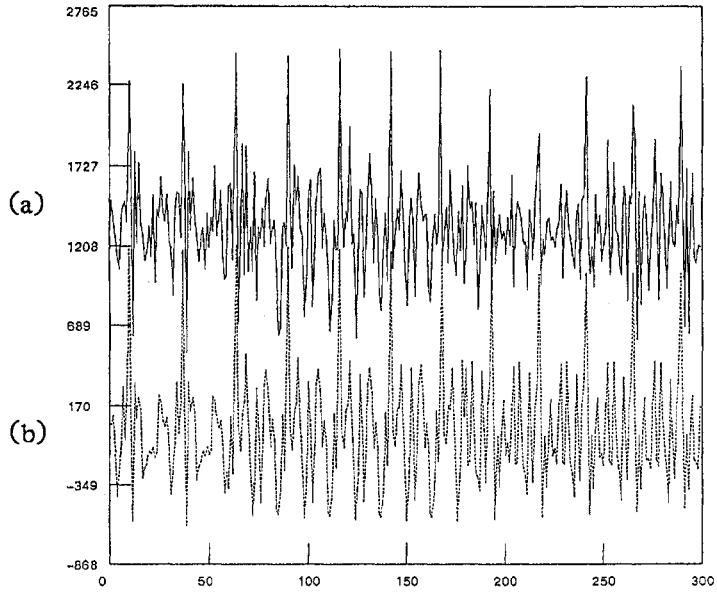

Figure 2 The waveforms of LPC excitation (a) Original (b) Reconstructed