A LOW RATE SINUSOIDAL SPEECH CODER

Mike Wu[♣] & W. H. Holmes[♥]

[♣]Motorola Australian Research Center
(formerly The University of New South Wales)

[♥]School of Electrical Engineering
The University of New South Wales

ABSTRACT - This paper reports on a study of low bit rate speech coding based on the sinusoidal model. Several techniques are employed to reduce the coder's bit rate and increase the processing speed. Firstly, the Hilbert transform is used to estimate the system phase response. Secondly, a bilinear transform is used to warp the spectrum to exploit the auditory characteristics of the human ear when coding at very low rates. Thirdly, a method to estimate the coarse pitch for the SEEVOC model is introduced and a method to perform the pitch correction is presented. Finally, a simplified birth-and-death algorithm is presented. The simulation results show that the reconstructed speech is of good quality at a bit rate of 4800 b/s, and is still intelligible and speaker recognizable at 1200 b/s.

1. INTRODUCTION

McAulay and Quatieri (1986, 1992) have developed a speech production system based on the sinusoidal model illustrated in figure (1), passing the excitation e(n) of a sum of sine waves through a linear time-varying filter, resulting in a sinusoidal representation for the speech waveform s(n),

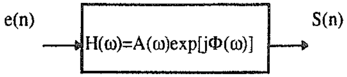

Figure (1) Sinusoidal speech production model

where

$$e(n)=\sum_{l=1}^{L}\exp\{j[\omega_l(n-n_0)]\}$$

$$s(n)=\sum_{l=1}^{L}A_l\exp\{j[\omega_l(n-n_0)+\phi_l]\} \qquad (1)$$

where A_l and ϕ_l represent the amplitude and phase of each sine wave component associated with the frequency ω_l, L is the number of sine waves, and n_0 is called onset time. It has been shown that equation (1) can also be simplified to a harmonic model. That is, the excitation frequency ω_l can be derived from the pitch frequency ω_0 and the voicing probability P_v, as shown in figure (2) (McAulay *et al.* 1986, 1992). The speech bandwidth is divided into two parts: voiced and unvoiced. The cutoff frequency is determined by the voicing probability P_v. In the voiced region the excitation frequencies are a set of pitch harmonics. In the unvoiced region the excitation frequencies are non-harmonics with 100 Hz intervals. Since the speech is represented by pitch frequency ω_0, system amplitude response A(ω) and phase response $\phi(\omega)$ as well as onset time n_0, it is desired that these parameters are estimated correctly and coded efficiently, using as few bits as possible. Also, in order to implement the coding in real time, the algorithms should be simplified. All the following analyses are based on a frame length with M=265 (about 33 ms), short time Fourier transform length N=1024, sampling rate Fs=8000 Hz. Hamming windowing is applied to the sampled speech data.

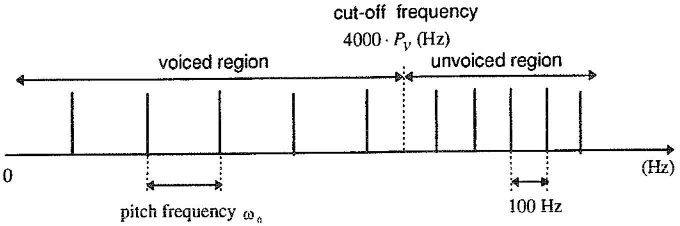

Figure (2) The exciation frequencies of the harmonic model

2. CODING SYSTEM RESPONSE H(ω)

2.1. Estimation of ϕ(ω)

By using the short time Fourier transform and the Spectral Envelope Estimation Vocoder (SEEVOC), the system amplitude response A(ω) can be obtained (McAulay et al. 1992, Paul 1981). If the system is assumed to be minimal phase, only the amplitude response need be encoded and transmitted, and the phase response ϕ(ω) can be recovered from A(ω), since ln [A(ω)] and ϕ(ω) are a Hilbert transform pair. In our low bit rate speech coding system a fast algorithm that uses the FFT was employed, as follows:

$$y(k) = FFT\{\ln[A(k)]\}, \quad k = 0,1,\dots N-1$$

$$z(k) = \begin{cases} 2 \cdot y[k] & \text{if } k < \frac{N}{2} \\ 0 & \text{if } k \geq \frac{N}{2} \end{cases}$$

$$\phi(k) = \mathrm{Im}\{IFFT[z(k)]\}$$

where A(k) and ϕ(k) are sampled A(ω) and ϕ(ω) respectively and Im denotes the imaginary part.

2.2. Coding the amplitude envelope

A main question is how to code the amplitude envelope at low bit rate. In a cepstral model, a small number (about 28) of cepstral coefficients is used to represent the amplitude envelope. Since the cepstral coefficients are not correlated and have a large dynamic range, it is difficult to code them. To avoid this problem, the cosine transform was applied to these cepstral coefficients (McAulay 1992), resulting in a series of what is called channel gains which have a small dynamic range and are highly correlated. The cosine transform pair used in this coding algorithm is as follows (Shenoi 1995):

$$g(k) = \sum_{n=0}^{R-1} c(n) \cos\left(\frac{2k+1}{2R}\pi\, n\right), \quad k = 0,1,\dots,R-1$$

$$c(0) = \frac{1}{R}\sum_{k=0}^{R-1} g(k)$$

$$c(n) = \frac{2}{R}\sum_{k=0}^{R-1} g(k) \cos\left(\frac{2k+1}{2R}\pi\, n\right), \quad n = 1,2,\dots,R-1$$

where g(k) (k=0, 1,..., R-1) are the channel gains, c(n) (n=0, 1,..., R-1) are the cepstral coefficients and R is the number of cepstral coefficients or channel gains used.

In order to provide a more efficient and perceptually valid allocation of the available bits, a spectral warping technique is employed in our coding system which samples more points of the amplitude envelope at lower frequencies than at higher frequencies. A bilinear transform (Fant, 1973) is applied to warp the spectrum before doing the cepstral transform. The bilinear transform can give a good approximation to the mel frequency scale, which is based on subjective pitch evaluations, and is as follows:

$$e^{-j\breve{\omega}} = \frac{e^{-j\omega} - \alpha}{1 - \alpha\, e^{-j\omega}}$$

where ω is the original frequency and $\breve{\omega}$ is the warped frequency, as shown in figure (3).

For a sampling rate of 8000 Hz and a bit rate of 1200 b/s, α should be in the range of 0.3 to 0.5. Our simulation results show that if the mel frequency warping is not applied when the bit rate is low (e.g. 1200 b/s), the reconstructed speech is of poor quality, but that it improves with warping. Figure (4) shows the recovered amplitude envelopes with and without the frequency warping technique, where the envelope is represented by 28 cosine-transformed cepstral coefficients (channel gains). It can be seen that at low frequencies the envelope is recovered better with frequency warping than without warping. The first channel gain is assigned 3 bits and the remainder are each assigned 1 bit, using a DPCM scheme .

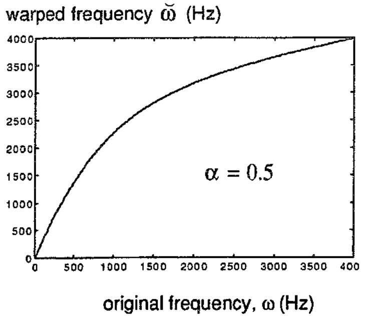

Figure (3a) Bilinear transform frequency warping

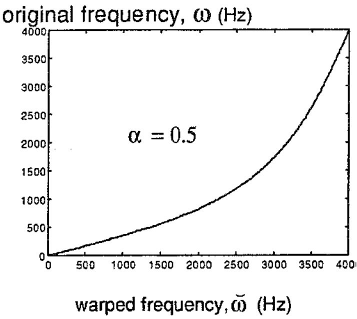

Figure (3b) Frequency dewarping

3. PITCH ESTIMATION AND VOICING DETECTION

3.1. Hybrid pitch detector

McAulay and Quatieri also developed a pitch estimation and voicing detection method based on the sinusoidal speech model that is inherently unambiguous, with less pitch doubling or halving problems than most other methods (McAulay et al. 1990). The MSE is the criterion used for the pitch estimation; that is, the method seeks the best ω_0 that minimizes the mean squared error

$$\varepsilon\left(\omega_0,\phi\right) = \frac{1}{N+1}\sum_{n=0}^{M-1} |s(n) - \hat{s}(n;\omega_0,\phi)|^2$$

where s(n) is the sinusoidal representation of speech and $\hat{s}(n;\omega_0,\phi)$ is the harmonic representation of speech. This will result in a likelihood function. An example is shown in figure (5), where the position of the maximum value corresponds to the pitch period and the maximum value can be used to calculate the voicing probability. However the computation load would be very high. A conventional pitch detector can be used to obtain a set of pitch candidates, which might consist of real, double and halved pitches, then a sinusoidal model pitch detector is used to select the best one from these candidates. The autocorrelation pitch detector can be used to generate the pitch candidates.

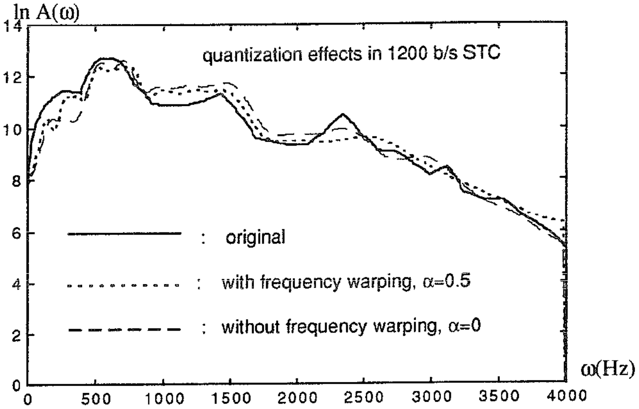

Figure (4) SEEVOC envelopes restored from 1.2 kb/s STC, with and without using frequency warping technique, compared with the unquantized one

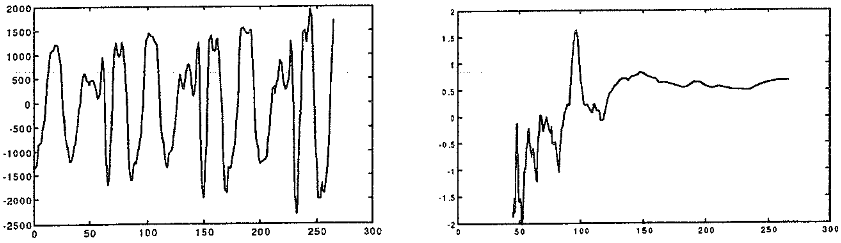

Figure (5a) Speech waveform

Figure (5b) Likelihood function for pitch estimation

3.2. Coarse pitch estimation and pitch correction

The SEEVOC model is crutial in the sinusoidal pitch detection, but the SEEVOC model itself needs a coarse estimate of pitch. Usually, the coarse pitch used in current frame can be the pitch estimated by the sinusoidal pitch detector in the previous frame. The sinusoidal pitch detector works well, provided that the frame length is adaptive to the pitch period, about 2.5 times the average pitch period. However, the bit rate of the STC is set to be fixed in this thesis, so is the frame length. Since the pitch can vary over a wide range, it is impossible to guarantee that the fixed frame length is always greater than 2.5 times the average pitch period. The pitch doubling problem often occurs when the frame length is less than 2.5 times the average pitch period. Besides, the pitch detector keeps working even in unvoiced or silent frames, where the concept of pitch is meaningless. A new method proposed to solve these problems is to keep track of both pitch frequency and voicing probability and then perform some smoothing operation on them.

Let $\omega_0^{(i)}$ and $P_v^{(i)}$ denote the pitch frequency and voicing probability in frame i, respectively, as estimated by the sinusoidal pitch detector. Let $\overline{\omega}_0^{(i)}$ denote the coarse pitch used for the SEEVOC envelope estimation in frame i. To obtain the coarse pitch, $\overline{\omega}_0^{(i+1)}$, used for the SEEVOC envelope estimation in frame i+1, postulate the following function,

$$\overline{\omega}_0^{(i+1)} = (1 - P_v^{(i)} \cdot P_v^{(i)}) \cdot \overline{\omega}_0^{(i)} + P_v^{(i)} \cdot P_v^{(i)} \cdot \omega_0^{(i)} \tag{2}$$

Equation (2) uses the voicing probability as a weighting factor to smooth the changes of coarse pitch. This weight was chosen after experiment and proves to be successful.

For the first frame a fixed coarse pitch, $\overline{\omega}_0^{(1)}$, is used for the SEEVOC envelope estimation. After a few frames, the coarse pitch will become accurate enough. This is because equation (2) can catch the right pitch and lock to it with relatively high probability. Therefore the SEEVOC envelope can be estimated correctly.

Incidentally, McAulay and Quatieri suggested that some other pitch detector might be used to obtain the coarse pitch for SEEVOC algorithm, but it is difficult to find such a pitch detector used for fixed frame length without pitch doubling or halving problems. The smoothing method described above works well in the simulation.

To solve the problem of pitch doubling or halving due to using fixed frame length, some simple rules are followed. If the relative change between $\omega_0^{(i)}$ and $\omega_0^{(i-1)}$ is greater than a set threshold, it is probable that pitch doubling or halving error has occurred, since the pitch usually will not change more than 40% between adjacent frames, according to the observations. If frame i is voiced, that is, if $P_v^{(i)}$ is large, a correction should be made — $\omega_0^{(i)}$ should be halved or doubled, whichever is closer to $\omega_0^{(i-1)}$. If frame i is unvoiced, that is, if $P_v^{(i)}$ is small, let $\omega_0^{(i)} = \omega_0^{(i-1)}$. In this way, pitch doubling or halving errors will be corrected.

4. RECONSTRUCTION OF SPEECH

4.1. Simplified birth-and-death algorithm

The concept of birth-and-death was introduced to match the parameters from frame to frame, making the reconstructed speech smooth (McAulay et al. 1986, 1992). The method of matching is to define sine wave tracks for frequencies that are successively "nearest neighbours". The matching algorithm itself is a rather tedious exercise. However, when the harmonic sinusoidal model is used, the algorithm can be greatly simplified, because with a harmonic model the frequencies are equally spaced. If the pitch $\omega_{0,k}$ (of frame k) is greater than $\omega_{0,k+1}$ (of frame k+1), each component in frame k can easily be matched to a component in frame k+1 – their frequencies are nearest neighbours. If $\omega_{0,k} < \omega_{0,k+1}$, reverse the matching direction; that is, let each component in frame k+1 find the matching component in frame k. The result of calculating such frequency tracks is illustrated in figure (6).

4.2 Observations on low bit rate sinusoidal coding

The quality of the reconstructed speech depends on the estimation of pitch, amplitude envelope and system phase response. The estimation of amplitude envelope is crucial since it not only recovers the system phase response at the decoder side, but also determines the success of the sinusoidal pitch detector.

In the sinusoidal coder, the SEEVOC method is used to estimate the amplitude envelope. An initial pitch estimate is needed for SEEVOC model, though it need not be very accurate. However, if the initial pitch estimate is too inaccurate, the envelope estimation will fail, so are pitch detection and phase recovery. The pitch correction scheme and the initial pitch estimation method presented in this paper have been proved successful in our low bit rate coder, especially in the case of fixed size frame.

In our coder, a simple linear interpolation is used to obtain amplitude envelope then a simple DPCM algorithm is used to code it (via channel gain.) According to our simulation results, when the coding bit rate is 4800 b/s, the reconstructed speech is of good quality; when the coding rate decreases to 1200 b/s, the reconstructed speech is still intelligible and speaker recognizab1e. At higher coding bit rates, since the envelope can be transmitted accurately, it is not necessary to perform spectral warping.

To obtain a better estimate of the envelope, Cheetham *et al.* (1995) suggest a optimization technique which optimizes the interpolation between the SEEVOC peaks and makes compensation for the inaccuracy of minimal phase assumption. At the present, the algorithm itself is computationally complex. Besides, in the low bit rate coding, the number of available bits is very limited therefore the envelope received at the decoder side will have some distortion. A better coding scheme for the envelope is also important. This work is currently under investigation.

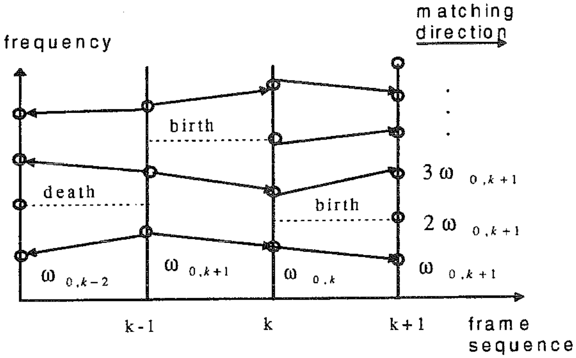

Figure (6) Frequency tracks in harmonic model

REFERENCES

Cheetham, B.M.G. & Sun, X.Q. & Wong, W.T.K. (1995) "Spectral envelope estimation for low bit-rate sinusoidal speech coders", *Eurospeech*, September 1995, pp693-696.

Fant, G. (1973) *Speech sound and features*, Cambridge: MIT Press.

McAulay, R.J & Quatieri, T.F. (1986) "Speech analysis-synthesis based on a sinusoidal representation", *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-34, pp. 744-754.

McAulay, R.J. & Quatieri, T.F. (1990) "Pitch estimation and voicing detection based on a sinusoidal speech model", *Proc. IEEE 1990 Int. Conf. Acoust., Speech and Signal Processing*, Albuquerque, pp. 249-252.

McAulay, R.J. & Quatieri, T.F. (1992) "Low rate speech coding based on a sinusoidal model", Chapter 1.6, pp. 165-208, in *Advances in Speech Signal Processing*, S. Furui and M.M. Sondhi (Eds.), Marcel Dekker, New York.

Paul, D.B. (1981) "The spectral envelope estimation vocoder", *IEEE Trans. Acoust., Speech and Signal Process.*, ASSP-29, pp. 786-794.

Shenoi, K. (1995) *Digital Signal Processing in Telecommunications*, Prentice Hall.