A HYBRID SUB-BAND SINUSOIDAL CODING SCHEME

M.S. Ho, D.J. Molyneux, B.M.G. Cheetham
Department of Computer Science, University of Manchester,
Oxford Road, Manchester M13 9PL, UK.
Email: msho@cs.man.ac.uk

ABSTRACT

This paper describes a hybrid sub-band speech coding scheme based on sinusoidal coding and CELP. Purely voiced speech is encoded using sinusoidal coding techniques and phase information is selectively transmitted. For mixed and unvoiced speech, the lower band is processed by sinusoidal coding algorithms while the upper band is encoded using CELP. To accommodate the extra bandwidth required by the encoded CELP parameters, the phase information is disregarded. The proposed coder is enhanced by sub-band discrete all-pole modeling and a voicing detection technique based on an analysis-by-synthesis approach. An efficient adaptive spectral shaping technique based on bandwidth widening in the LSP domain is employed. The proposed technique is capable of producing high quality speech at 4.1 kbit/s.

1. INTRODUCTION

Sinusoidal models have been shown to be able to represent voiced speech accurately. However, when applied to low bit-rate coding of speech, an inadequacy of sinusoidal models often manifests itself in the quality of the mixed or unvoiced speech produced, leading to objectionable artifacts or unnatural sounding unvoiced speech. Much research effort has been made to address this issue. Macon and Clements introduced phase randomization to the sinusoidal model of unvoiced speech [1]. Recently, Stachurski and McCree designed a hybrid MELP/CELP coder [2] while Nishiguchi, et al [3], and Nakhai and Marvasti [4], proposed hybrid speech coding schemes based on harmonic coding and CELP. These hybrid coders are similar in that CELP is used to encode unvoiced speech frames. The motivation for adopting this approach is that CELP is essentially a waveform matching algorithm and is therefore very effective at representing irregular or noise-like features of speech. It seems reasonable and beneficial to combine sinusoidal coding and CELP in order to achieve high quality speech at very low bit-rates.

However, a common problem often associated with hybrid coding schemes is that frequent switching from one mode to another may produce undesirable artifacts. Moreover, a hard voicing decision will be necessary to select the algorithm used to encode each frame. This paper adopts a strategy in which both sinusoidal coding and CELP techniques run in parallel. Each algorithm handles a sub-band of the speech. In the traditional sinusoidal transform coding (STC) [5], a voicing cutoff frequency \( f_v \) splits the 0 to 4kHz frequency range into a voiced and an unvoiced band, the latter being parameterised according to a model consisting of sinusoids with random phase. In our approach, the lower band is modeled using sinusoidal coding techniques while the upper band is encoded using CELP.

The quality of speech produced by this hybrid coding scheme to a considerable extent depends on the accuracy of \( f_v \). An analysis-by-synthesis voicing detection algorithm [6] has been developed to obtain an objectively more accurate \( f_v \) and thus a more natural synthesised speech. In this procedure, the mean square difference between the magnitude spectra of sinusoidally modeled speech using (a) the measured sinusoids and (b) harmonically related sinusoids with “pitch” frequency \( f_0 \) is minimised over a frequency range normally extending up to 4 kHz. This algorithm also provides a basis for sub-band discrete all-pole modeling (SB-DAP) of speech spectra. The spectral envelope produced by SB-DAP is adaptively regularised using a novel bandwidth widening technique operating in the LSP domain [7].

For strongly voiced speech, sinusoidal coding techniques alone are capable of highly detailed pseudo-periodic speech waveforms. Where \( f_v \) is close to 1 and there is no necessity to encode an unvoiced upper band, there is bit-rate capacity available for preserving the fidelity of the voiced speech waveform by including phase information. Traditional sinusoidal techniques model the phase spectrum on the assumption that the transfer function of the vocal tract model is minimum phase. The phase spectrum is therefore derived from the spectral envelope using a Hilbert transform. Although capable of producing good quality speech, the minimum phase assumption is often disputable, as the opening and closing of the glottis produces a pressure waveform that resembles a Rosenberg pulse which is clearly not minimum phase [8]. By transmitting the more perceptually significant phase information, i.e. the phases of the low frequency harmonics, it is possible to reproduce voiced speech closer to the original. In our approach, the measured phases of the first six sinusoids are transmitted. At the synthesiser, the linear phase component is estimated from the decoded measured phase [9]. The high frequency phase spectrum is regenerated from the estimated linear phase component and the minimum phase spectrum. The voiced speech frame is reproduced entirely by a bank of oscillators. The proposed hybrid sub-band coding scheme is capable of producing high quality speech at 4.1 kbit/s.
2. OVERVIEW OF THE CODING SCHEME

2.1 Encoder

Figure 1(a) shows the structure of a coder based on the approach in this paper. Windowed speech is analysed pitch-synchronously using the sinusoidal modeling techniques [5] at an update rate of 15 ms. Prior to pitch estimation, linear pitch-period variations in the speech frame is removed by means of time-warping [10]. The voicing cutoff frequency $f_v$ is estimated by applying an analysis-by-synthesis voicing detection algorithm [6] to the warped speech. Sinusoidal magnitudes are represented by a 10th order all-pole filter optimised using sub-band DAP. The filter coefficients are converted to LSP and the spectral envelope is then processed by an adaptive bandwidth widening process in the LSP domain. Having determined the voicing and magnitude information, the encoder then splits into 2 modes according to the voicing decision. When $f_v$ is 1 (i.e. strongly voiced), the measured phases of the first six sinusoids are vector-quantised. When $f_v$ is smaller than 1 (i.e. unvoiced or mixed speech), the Fourier spectrum below $f_v$ is zeroed followed by an inverse FFT operation. The unvoiced speech thus derived is split into two 7.5 ms sub-frames. Each sub-frame is then encoded using CELP algorithm with a 8-bit Gaussian codebook.

2.2 Decoder

Figure 1(b) shows the block diagram of the decoder. The framing at the synthesiser is divided into two 7.5 ms sub-frames. The parameters of each sub-frame are obtained by interpolating the pitch, voicing, gain and LSP between two update points. Strongly voiced speech is reconstructed by a bank of oscillators across the entire speech band using the following relationship:

$$ s(n) = \sum_{l=0}^{L} \hat{A}_l \exp(j(n-n_0)\omega_0 + \Phi(\omega_0) + \beta \pi) $$

where $L$ is the total number of harmonics up to 4 kHz, $\hat{A}_l$ is the decoded sinewave amplitudes and $\omega_0$ denotes the fundamental frequency in radian per second. The minimum phase spectrum $\Phi(\omega)$ is derived using a Hilbert transformer from the spectral envelope. The ambiguity bit $\beta$ accounts for the sign of the input speech waveform and $n_0$ is the onset time at which sinusoids come into synchronization. Since $n_0$ and $\beta$ are not transmitted, they are estimated from the decoded measured phases at the decoder [9]. It can be shown that an "optimum" onset time can be derived by maximizing $\rho(n_0)$:

$$ \rho(n_0) = \sum_{l=1}^{L} \hat{A}_l^2 \cos(\hat{\theta}_l + n_0\omega_0 - \Phi(\omega_0)) $$

Figure 1: Schematic diagram of (a) encoder (b) decoder
where $\hat{\Theta}_i$ is the decoded phase.

For mixed or unvoiced speech, only sinusoids up to $f_v$ are generated. The speech component above $f_v$ is obtained at the output of an LP synthesis filter. The excitation to the filter is provided by a CELP Gaussian codebook and gain. The voiced and unvoiced components are summed to form the mixed/unvoiced speech frame. The reconstructed speech frame can be represented as

$$\hat{s}(n) = \sum_{i=0}^{M} \hat{A}_i \exp(jf_v(n - n_{ic})(\omega_i + \Phi(\omega_i))) + g(n)\hat{v}(n) - \sum_{i=0}^{\hat{N}} a_i\hat{s}(n-i)$$

where $M = \text{int}[f_v/f_q]$, $g(n)$ and $\hat{v}(n)$ denotes CELP gain and codeword respectively, and $a_i$ are the coefficients of an LP filter of order $p$. To ensure continuity at the frame boundaries, the overlap-add synthesis procedure is carried out on each sub-frame using a triangular window.

3. ADAPTIVE SPECTRAL ESTIMATION

Traditional STC [5] models spectral magnitudes by fitting an all-pole model to a cubic spline envelope of spectral peaks. In CELP algorithms, conventional LP spectral estimation is generally used. In sub-band hybrid speech coding schemes, it may be beneficial to represent the spectral information in each sub-band by a different mechanism. It is desirable to derive spectral envelopes that represent sinewave magnitudes accurately for voiced speech components, and also allow a reasonable LPC synthesis filter to be deduced for mixed and unvoiced speech components as required in CELP. DAP [11] is an alternative to more traditional forms of LP analysis as a method of obtaining an all-pole representation of the short-term spectral envelope of a segment of voiced speech. It is based on the minimisation of a discrete version of the Itakura-Saito (IS) distance between the all-pole spectral envelope sampled at discrete frequencies and spectral amplitudes derived from the STFT spectrum usually by peak-picking. DAP may be expected to be effective, and better than traditional LP, for voiced regions of a spectrum which are sampled at the pitch-harmonics. However, the same process is unsuitable for parts of the spectrum considered unvoiced because each DFT spectrum will provide, in theory, a statistical estimate of the true power spectral density and will exhibit random frame to frame variation. Peak-picking as used in the voiced band below $f_v$ would produce highly variable non-all-pole shapes even when the unvoiced speech band conforms well to an all pole model.

We believe it is therefore more appropriate to perform spectral estimation equivalent to traditional LP for unvoiced regions of the spectrum to produce a continuous frequency fit. It is known that traditional LP works well in theory and in practice when we can consider the signal to be the result of exciting an all-pole transfer function with white noise rather than a periodic signal. DAP can then be reserved for the voiced spectral region below the $f_v$. In order to achieve this aim, we propose that the unvoiced part of the traditional LP spectral envelope replaces the corresponding unvoiced part of the true speech spectral envelope when the DAP algorithm described above is applied to speech spectra split by $f_v$. This incurs no additional computational cost since traditional LP analysis is always the first iteration of the DAP algorithm. The voiced and modified unvoiced regions of the magnitude spectrum are both sampled at multiples of the pitch-frequency as measured for the voiced region. Some adjustment to the spectral peaks is made in the voiced region as in the traditional SEEVO algorithm however, there will be no harmonically related peaks in the modified unvoiced part of the spectrum. The same spacing between discrete spectral samples is used in both the unvoiced and the voiced bands to achieve even spectral weighting across the whole speech spectrum. The samples of the LP envelope above the voicing transition frequency are guaranteed to correspond to an all-pole model therefore the adaptive sub-band DAP algorithm should have little difficulty in producing an accurate spectral fit to this envelope.

Due to the effects of lack of pitch stationarity and slightly over resonant peaks due to the model order being too high on occasions, audible distortions can occur from time to time in the form of short tonal bursts. This means that the over resonant poles must be reduced. Fixed bandwidth widening affects the bandwidths of all resonances equally. It is probable that some formant resonances will require more bandwidth widening than others depending on the closeness of the interpolated pitch harmonic frequencies to their centre frequencies. In order to take this into consideration an adaptive bandwidth widening scheme [7] is applied to the DAP modeling technique. In this approach, the closeness of the interpolated pitch harmonics to the centre frequencies of DAP estimated resonances is considered when calculating the degree of bandwidth widening. The formant frequencies are estimated from identifiable patterns of closely spaced LSP coefficients, where each formant can often be associated with a pair or triple set of LSP coefficients [12]. The closeness of poles to unit circle is estimated from sets of LSP’s, where poles too close to unit circle produce over-resonant formants. These can be widened by pulling the corresponding LSP’s further apart from each other.

4. PHASE CODING

Strongly voiced speech is synthesised by summing harmonic sinewaves. To achieve high quality speech, the phases must be quantised every 12.5ms using 5 bits for each phase. At a sampling rate of 8 kHz, the total number $N$ of harmonics varies from 10 to 80 depending on the fundamental frequency, requiring a very high bandwidth which is not feasible in low-rate applications. However, it has been observed that for strongly voiced speech, it is possible to derive the details of the voicing information and to preserve the temporal structure of the speech waveform using the measured phases of the first six harmonics. In our approach, the first six phases are encoded by a 2-dimension vector quantiser using a 7-bit codebook [13] every 15ms. To improve the temporal resolution, the phases of the inner sub-frame are predicted by a cubic phase interpolation [14] function derived from the decoded phases of two adjacent outer frames. Since only the first six measured phases are transmitted, the remaining high-frequency phase spectrum is regenerated from the minimum phase spectrum and an onset time estimated from the decoded phases using equation 2.

5. BIT ALLOCATION

The bit allocation for the hybrid coder is given in Table 1. In both voiced and mixed/unvoiced mode, the 10th order SB-DAP
spectral envelope is represented by 10 LSPs which are quantised using split vector quantisation at 24 bit per frame. The pitch period is uniformly quantised with 7 bits. The voicing cutoff frequency $f_v$ is quantised using a 3-bit scalar quantiser. Prior to quantisation, $f_v$ is smoothed by a simple averaging operation with a look ahead and a look backward of 15 ms. Frame energy is predicted using a simple first order predictor and quantised non-linearly with 6 bits. In voiced mode, three 7-bit phase codebook indices are transmitted together, amounting to 21 bits per 15 ms. In mixed/unvoiced mode, 11 bits were used to encode the CELP parameters of a 7.5 ms sub-frame. 3 bits of the CELP bit stream are allocated to the codebook gain. A total of 22 bits is therefore required to encode the CELP parameters every 15 ms. An unused bit in the voiced mode is reserved for future improvements.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>SC</th>
<th>SC/CELP</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSP</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Energy</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Pitch</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Voicing cutoff</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Phase error codebooks</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>CELP Gaussian codebooks</td>
<td>-</td>
<td>16</td>
</tr>
<tr>
<td>CELP gains</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>Reserved</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>TOTAL</td>
<td>62</td>
<td>62</td>
</tr>
</tbody>
</table>

*Table 1: Bit allocation*

6. TESTING

Informal listening tests on speech of various pitch range spoken by male and female speakers suggests that the 4.1 kbit/s hybrid coder is capable of producing more natural sounding speech than codecs of similar bit-rates such as FS1016 CELP and the Inmarsat IMBE. The waveform shape tends to be better preserved than with the minimum phase assumption, as illustrated in Figure 2. The minimum phase assumption tends to exaggerate the peak to mean ratio which can have a number of disadvantages. It is planned that more sophisticated quantisation techniques may be implemented to lower the bit-rate further.

7. CONCLUSION

We have presented a high quality 4.1 kbit/s hybrid speech coding scheme based on sinusoidal coding and CELP. Speech is split into 2 sub-bands, the lower band is processed by sinusoidal coding algorithms while the upper band is encoded using CELP. The coder is enhanced by a novel sub-band DAP technique, an improved voicing detection and phase coding.

8. ACKNOWLEDGMENTS

This work was carried out at the University of Manchester with support from the UK Engineering and Physical Sciences Research Council, British Telecom, the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom and the University of Manchester.

9. REFERENCE


