VARIABLE BIT-RATE SINUSOIDAL TRANSFORM CODING USING VARIABLE ORDER SPECTRAL ESTIMATION

N. Li, D.J. Molyneux, M.S. Ho, B.M.G. Cheetham

Department of Computer Science, University of Manchester,
Oxford Road, Manchester M13 9PL, U.K.
E-mail: lin@cs.man.ac.uk

ABSTRACT

Sinusoidal transform coding (STC) is known to be capable of producing good communication quality speech coded at bit-rates below 4kb/s. Discrete all-pole modelling (DAP) is an alternative spectral estimation method which can be more accurate than the conventional linear prediction (LP) analysis normally used by STC. In the quest to achieve the highest possible speech quality at lower and lower average bit-rates in variable bit-rate coding schemes, more and more effort must be made to investigate ways of varying the number of parameters according to the characteristics of each speech frame. This paper considers the advantage to be gained by varying the all-pole model order according to the discrete Itakura-Saito (IS) distance measure used in DAP. A significant reduction is achieved in the average number of parameters to be quantised compared to the fixed order model while the speech quality remains the same.

1. INTRODUCTION

Third generation systems will open a new era of mobile communications with a combination of wireless speech communication, Internet and multimedia services. Mobile users will enjoy the same access to information via the Internet as fixed users. Speech channels may therefore be competing for available bandwidth with an increasing number of other data streams. Also, it will take some time for broadband third generation systems to become universally available. Hence, as subscribers become accustomed to the increasing functionality of third generation mobile telephony, it may be anticipated that demands for enhanced communication features will be created in areas served only by lower bandwidth mobile systems. This means there will be a demand for speech and visual information to be transmitted over bandwidths previously used for speech only, creating a continuing need for high quality low bit-rate speech coding.

The application of low bit-rate speech coders to meet the demands of Internet telephony and mobile radio has led speech coder designers to introduce flexibility into the speech model used, with frame to frame variations based on the changing characteristics of the source signal. During a normal telephone conversation, the amount and the content of information in each frame will be variable. Some frames will contain only background noise if one of the participants is not actually speaking. Even when he or she is speaking, there will still be some frames, containing unvoiced speech for example, that are easier to digitise than others. Varying the bit-rate according to the speech being digitised is referred as “source control”. This paper is concerned with source-controlled speech coding using sinusoidal transform coding (STC) combined with discrete all-pole modelling (DAP). The aim is to vary the number of parameters used to characterise the short-term spectrum of a frame according to the “discrete Itakura-Saito (I.S.)” distance measure between the resulting all-pole model and the true spectral envelope. We concentrate on active speech frames assuming that a voice activity detector (VAD) has eliminated any frames containing only background noise.

2. SPEECH CODING

Over the past decade, code-excited linear predictive (CELP) coding has been widely adopted for many standards but at bit-rates below 4kbit/s, the speech quality tends to become synthetic sounding [1]. This can be attributed to the fact that with reduced bit-rate, accurate coding of the excitation signal becomes more difficult, and the spectral modelling becomes more critical. With the continuing advance of research in low bit-rate speech coding, sinusoidal transform coding (STC) has proved to be one of the most promising techniques for coding speech of good quality and intelligibility at low bit-rates.

STC is based on a sinusoidal model of speech [1], whose parameters are the amplitudes, frequencies and phases derived from a high resolution short-term Fourier transform performed at intervals of 20 to 30 ms. A “voicing probability” frequency $f_v$ splits the 0 to 4kHz frequency range into a pseudo-periodic lower band and a non-periodic upper band. The sinusoidal model parameters are efficiently encoded using a fundamental frequency estimate, the voicing probability frequency and an all-pole approximation to the short-term spectral envelope represented by line spectral pair (LSP) coefficients. At the decoder, the spectral envelope is sampled at the excitation frequency harmonics below $f_v$ Hz and at fixed frequencies above $f_v$, Hz to obtain the required sinusoidal amplitudes and frequencies. The phase values below $f_v$ are calculated by sampling a phase spectrum obtained from the spectral envelope via a Hilbert transform. Above $f_v$, a random phase spectrum is synthesised. Speech is reconstructed as a linear combination of sinusoids with time-varying amplitudes, frequencies and phases thus derived.

For each speech frame, traditional STC fits an all-pole model to a spectral envelope obtained by selecting spectral peaks using a SEEVOC [2] algorithm, and fitting cubic splines to interpolate between the peaks. Even for fully voiced speech, this envelope cannot be expected to conform exactly to an all-pole transfer function, particularly between the pitch harmonics. The discrepancy becomes especially significant when vocal tract resonances lie between adjacent harmonics. Such resonances are flattened by the cubic spline interpolation and any variation in
pitch will cause the interpolated envelope to change, even for fixed vocal tract resonances. LP analysis, as applied by STC to derive its all-pole model, minimises the continuous Itakura-Saito distance measure between the model’s magnitude spectrum and the cubic spline interpolated spectrum. In addition to well known inaccuracies in LP analysis, unnatural frame to frame variation occurs particularly when the cubic spline envelope is fitted to spectra where the pitch has varied over the speech frame. At the decoder distortion may occur in the phase spectrum, as it is calculated from the estimated envelope via a Hilbert transform. McAulay and Quatieri [1] have reported that the minimum order of all-pole model required to achieve speech quality equivalent to that obtained using the true unquantised cubic spline envelope are 22. Proposed STC algorithms [1] employ 14th order all-pole models with perception-based spectral warping. A 10th order all-pole model of the cubic spline envelope is reported to produce synthetic speech that is mechanical and buzzy [1].

The potential for representing the spectral envelopes of voiced speech frames using an all-pole transfer function derived using a discrete all-pole modeling (DAP) [3] based approach has been investigated [2, 4]. DAP accurately models spectra, which conform well to an all-pole model, but is not immediately appropriate for modeling non-periodic (unvoiced) regions of speech spectra. To adapt DAP to STC, an alternative method of modeling the unvoiced upper frequency band has been developed [5].

Speech may be categorised as either voiced, unvoiced or mixed depending upon the degree of periodicity observed in the excitation signal. In the case of fully voiced sounds the vocal tract is predominantly resonant in nature and may be expected to conform well to an all-pole model with each complex-conjugate pole pair representing a vocal tract resonance. In principle, an all-pole model of order ten, capable of representing up to five different resonances, should be capable of achieving a spectral representation of sufficient accuracy for good quality sinusoidally modeled speech. Fully unvoiced speech is considered to have a more random nature with a random spectral distribution which will vary significantly frame to frame. The exact shape of the short term power spectral envelope, as measured for each unvoiced frame, is perceptually less important than for voiced speech, and a much lower order all-pole model will often suffice. For mixed frames a compromise between the above conditions is necessary and only the voiced lower frequency band will require accurate parameterisation. A clear advantage can be achieved if the speech coder can determine to what extent frames require higher order vocal tract models for strongly resonant voiced bands and allocate an appropriate order to the model. In the algorithm presented in this paper, the order of the all-pole model varies frame to frame to maintain an acceptable discrete L.S. measure between the synthesised and original spectral envelopes with minimum order.

3. DISCRETE ALL-POLE MODELING (DAP) APPLIED TO STC

The DAP technique [3] aims to determine the parameters $a_m, a_p, a_2, ..., a_p$, of a $p^{th}$ order all-pole transfer function, $1/A_f(\omega)$, such that the corresponding power spectrum $\hat{P}(\omega)$ is as close as possible to the power spectrum $P(\omega)$ of the speech frame at a number of discrete frequencies which are normally harmonics of the excitation frequency for the voiced band and appropriate fixed frequencies for the unvoiced band. The measure of closeness is a discretised version of the Itakura-Saito(IS) distance measure [3] which may be written as:

$$E = \frac{1}{p} \sum_{m=1}^{L} \left( \frac{P_m}{\hat{P}_m(\omega)} - \ln \left( \frac{P_m}{\hat{P}_m(\omega)} \right) - 1 \right)$$

where $\omega_m$ for $m = 1, 2, ..., L$. are the chosen discrete frequencies in the range $0$ to $2\pi$ radians/sample and $P_m = P(\omega_m)$ is the power spectral density of the speech at frequency $\omega_m$. In contrast to traditional LP analysis which minimizes the continuous IS distance measure between $P(\omega)$ and $\hat{P}(\omega)$ and is affected by the spectral shape of $P(\omega)$ between harmonics, DAP is only concerned with the spectral match at the chosen discrete frequencies. In comparison to the true vocal tract frequency-response, the measured spectrum $P(\omega)$ is effectively down-sampled by the vocal tract excitation and the result is a form of time-domain aliasing where the responses to consecutive pitch pulses run into each other. The aliasing occurs in natural speech where the number of pitch-harmonics is insufficient to accurately characterize the shape of the vocal tract frequency response. It affects the autocorrelation function and hence the all-pole model derived by traditional LP analysis. In minimizing the discrete IS measure, DAP matches to the aliased autocorrelation function of natural speech a similarly aliased autocorrelation function as would be obtained by down-sampling the power spectrum of the required all-pole model at the discrete frequencies, and applying an inverse Fourier transform.

DAP can produce all-pole models which accurately fit purely voiced spectra sampled at well-defined harmonic frequencies [4]. In order to apply DAP to STC, a means of modeling the spectrum above the voicing probability must be found. This part of the spectrum is considered unvoiced and is therefore not pseudo-periodic. A sub-band approach may therefore be used [5]. Firstily peaks are picked in the sub-band below the voicing probability as with traditional STC. For the upper frequency band, considered unvoiced, a set of amplitudes are calculated by sampling the traditional LP filter gain response (used as the initial DAP iteration) at the harmonics of the fundamental frequency as derived for the voiced lower frequency band. By sampling the LP envelope estimate of the unvoiced region a set of magnitudes is obtained which cause DAP to produce a reasonable spectral envelope shape. Once the discrete amplitudes have been identified in both frequency bands, the traditional DAP algorithm described above may be used.

The frequency $f_c$ plays an important role when DAP is employed in STC. A novel algorithm for determining $f_c$, based on a frequency-domain analysis-by-synthesis optimisation procedure, has been developed[6]. This produces a more reliable voicing probability frequency than traditional STC and
thus allows the spectral envelope for the whole frame to be more effectively and economically modeled at variable bit-rates.

4. DAP MODEL ORDER SELECTION

Increasing the order of the vocal tract model with traditional LP analysis increases the range of delays over which the autocorrelation function of the true speech spectrum and that of the all-pole spectrum are equal. Thus a better fit of the true speech power spectrum to that of the all-pole transfer function is achieved and the continuous IS distance measure will decrease monotonically with increasing model order. Similar behavior is to be expected from DAP. To demonstrate this, a segment of artificial speech with three poles was produced and, starting from order one, a discrete IS distance measure was derived for each DAP model order up to ten. To do this efficiently, instead of starting the DAP iteration procedure from the LP spectrum in each case, the lower order model from the previous stage was used as the starting point for the next stage. As expected, \( E_p \) reduces monotonically and becomes close to zero at order six as seen in Figure 1 (solid line). There is a clear leveling off of the curve at order six, though the reduction in \( E_p \) from order 2 to 3 is small.

\[
\begin{align*}
\text{Figure 1: } E_p \text{ as a function of } p \text{ for pure (solid line) and noisy (dashed line) artificial speech.}
\end{align*}
\]

When white Gaussian noise was added to the artificial speech to achieve a S/N ratio of 11dB, the leveling off of the curve became less sharply defined as shown in Figure 1 (dashed line).

With real speech, increasing the DAP model order generally improves the spectral envelope match, and in many cases, the improvement above a certain order also becomes insignificant. An improvement criterion is therefore applied to select the order \( M \) for each frame of speech; i.e. \( M \) is selected such that the reduction in \( E_p \) obtained by increasing the order, \( p \), above \( M \) is below a predetermined threshold. The minimum order is set to 4. The maximum order is ten because fixed 10th order DAP has been found to produce good quality speech comparable to 14th order conventional all-pole modeling [5]. The order selection procedure starts from the minimum order and increases this to the maximum, calculating \( E_p \) in each case. The difference between successive values of \( E_p \) and \( E_{10} \) is calculated. If the difference is below a certain threshold, which means that the improvement of the spectral envelope is not sufficient to justify an increase in order, then the lower order is selected. The value of the threshold was made dependent upon the voicing probability. For a voicing probability below 0.5, a threshold of 0.1 was chosen and for voicing probabilities above 0.5, a threshold was taken to be 0.01. These choices were found experimentally to be reasonable.

\[
\begin{align*}
\text{Figure 2: } E_p \text{ as a function of } p \text{ for a frame of real speech.}
\end{align*}
\]

The dependency of \( E_p \) on \( p \) for a frame of fully voiced speech is illustrated in Figure 2. In this case, there is no great decrease in \( E_p \) with increasing \( p \) from order 8. Hence, order 8 is selected producing the spectral envelope shown (dotted line) in Figure 3. This algorithm will obtain a bit-rate saving, with a moderate increase in computational intensity. Figure 4 presents the variation in model order selection for a short speech segment consisting of about 2 seconds of speech. In this experiment, only even orders were considered.

\[
\begin{align*}
\text{Figure 3: Spectral envelope for } 8^{th} \text{ (dotted line) and } 10^{th} \text{ (solid line) order DAP.}
\end{align*}
\]

\[
\begin{align*}
\text{Figure 4: Result of DAP model order selection}
\end{align*}
\]

We adopted the bit allocation scheme of the various variable model order codebooks from 6 to 10, presented in [7]. The
order uses a single 10-bit codebook. The distribution of model orders with selected bit configurations for a period of active speech (without silence frames) about 47 minutes long which consists of approximately 170,000 frames of 20ms is presented in Table 1.

Table 1 shows that the overall reduction in average bit-rate for the spectral parameters is 6.9 bits per frame, i.e. from 25 to 18.1, assuming the LSP parameters are quantised as specified above. This takes into account the mode information needed to specify the order.

<table>
<thead>
<tr>
<th>Model order</th>
<th>Probability (%)</th>
<th>DAP bits</th>
<th>Mode bits</th>
<th>Average bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>49.6</td>
<td>10</td>
<td>16</td>
<td>16.1</td>
</tr>
<tr>
<td>6</td>
<td>5.2</td>
<td>16</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>20.1</td>
<td>20</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>25.1</td>
<td>25</td>
<td>-</td>
<td>25</td>
</tr>
<tr>
<td>10(Fixed)</td>
<td>100</td>
<td>25</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1. Bit-rate comparison between variable rate and fixed rate DAP

Testing was carried out using both objective and subjective measurements. Firstly, log spectral distortion (LSD) measurements were obtained to test the accuracy of fit of the estimated spectral envelope. Secondly, mean opinion scores (MOS) were obtained to assess the speech quality and the distortion for examples of both male and female speech. In [5], LSD measurements and MOS tests indicate that the application of DAP to STC improves speech quality over what is achievable with 14th order all-pole modeling of the cubic spline envelope. Hence, the tests here are carried out between the fixed 10th order DAP and the variable order DAP. The results are given in Table 2 and show that variable order DAP has a similar average spectral distortion to that obtained for fixed 10th order DAP.

<table>
<thead>
<tr>
<th>Log Spectral Distortion</th>
<th>10th Order DAP</th>
<th>Variable Order DAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.64dB</td>
<td>1.55dB</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Spectral distortion for fixed & variable order DAP.

Informal subjective listening tests were carried out using the listening quality scale (bad=1; poor=2; fair=3; good=4; excellent=5) to give a Mean Opinion Score (MOS) style of measurement. The tests used two recordings of female speech and two recordings of male speech, each recording consisting of three sentences. Twenty subjects performed the tests and 240 votes were given in total. Table 3 presents the subjective test results.

<table>
<thead>
<tr>
<th>Gender</th>
<th>10th Order DAP</th>
<th>Variable Order DAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3.30</td>
<td>3.25</td>
</tr>
<tr>
<td>Female</td>
<td>3.59</td>
<td>3.65</td>
</tr>
</tbody>
</table>

Table 3. Subjective comparison between fixed & variable order DAP.

The listening test results in Table 2 indicate that variable order DAP is considered similar in quality to fixed 10th order DAP for examples of male and female speech. On average, variable order DAP does not appear to degrade the speech quality while decreasing the number of spectral parameters to be quantised.

5. CONCLUSION

The objective of this work was to investigate source controlled variable order spectral estimation for a variable bit-rate DAP-STC codec. The model order is determined by measuring the discrete I.S. distance between spectral envelopes. The result is a lowered average bit-rate for the parameter quantization with no perceivable reduction in speech quality.

6. ACKNOWLEDGEMENTS

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 United Kingdom and the University of Manchester.

7. REFERENCES