SEGMENTAL PROTOTYPE INTERPOLATION CODING

Professor C. S. Xydeas, Thomas M. Chapman
Speech and Image processing laboratory
Simon Engineering Building
The University of Manchester
Brunswick Street, Manchester, M13 9PL, UK

ABSTRACT

Current parametric speech coding schemes can achieve high communications quality speech at bit rates in the range of 2.4 to 1.5kbits/sec. Most schemes sample and quantise, at regular intervals, the "tracks in time" generated by the parameters of the speech production model. As a result, reconstructed "parameter tracks" do not evolve "smoothly" with time. Furthermore, no advantage is taken of the "linguistic event" nature of speech. In this paper, model parameter "time tracks" are split into non overlapping speech "event" related segments. These segment based *evolutions* of model parameters are then vector quantised to provide at the receiver a smooth and subjectively meaningful reconstruction. Thus the paper presents an application of this generic segmental speech model quantisation approach to a 1.5kbits/sec Prototype Interpolation Coding (PIC) system. Results indicate that the proposed methodology can almost halve the bit rate of this PIC system while preserving overall recovered speech quality.

1. INTRODUCTION

Recent low bit rate speech coding research has focused upon the application of Waveform Interpolation Coding [1, 2, 3, 4], Sinusoidal coding [5] and Multi Band Excitation [6] techniques on either the speech or the LPC residual signals. In each of these methods, model parameters are calculated, quantised and transmitted on a "frame by frame" basis. Although these models are sound, a fundamental disadvantage using conventional frame based quantisation is that the evolution with time of the model parameters is not "smooth", but is rather uneven, since quantisation distortion alters randomly in "direction" between coding frames.

In very low bit rate coding (<400 bits / sec), work has focused upon the quantisation of "phonetic" units / segments [7, 8] and successful methods for the isolation and classification of such units have been demonstrated. However the underlying speech production model used in such schemes is often relatively crude in comparison with those employed at relatively higher bit rate (1.5 - 2.4 kb/s) systems.. In addition, many of these systems are usually speaker and training corpus dependent.

This paper brings together the successful speech production models applied at higher rates, the ideas behind the decomposition of the speech signal into phonetic units and segmental quantisation [9, 11] techniques. This leads to a Segmental PIC system which operates with high communications quality speech at rates in the range 800 - 1000 bits / sec. Thus the proposed generic segmental speech model quantisation methodology combines the strengths of prototype interpolation speech synthesis models with the additional bit rate savings obtained through the application of "phonetic" based segment coding.

In the following sections, we firstly describe briefly the Manchester Pitch Synchronous PIC model which has already been used to produce high communications quality speech at 1500 bits / sec. Using this as a basis, a novel segmental quantisation scheme is then proposed which allows the model to be applied successfully at 800 - 1000 bits/sec while maintaining the speech quality of the 1.5kbits/sec Codec. System performance is discussed in the final section.

2. MANCHESTER PITCH SYNCHRONOUS (MPS) PROTOTYPE INTERPOLATION CODING

Figure 1 shows a block diagram of the encoding stage of the MPS-PIC model. Essentially, the coder consists of an LPC analysis stage, an integrated pitch estimation and voiced / unvoiced classification algorithm and a pitch synchronous encoding of the excitation signal. In addition, a "mixed voicing excitation" model [3] adds randomness, as required, to the harmonics of output voiced frames, in order to increase the naturalness of the recovered speech signal. The system operates on 20ms analysis frames. Briefly, the individual processes involved in the encoder are as follows [2]:

Process I

In this process the voicing status V^n and pitch estimate P^n of the current (n^{th}) input frame are determined. This is performed using a novel Voicing and Pitch estimation algorithm which jointly estimates the voicing nature and pitch of an input speech frame.

Process II

In this process, the Burg algorithm is used to calculate the coefficients $\{a_i^n\}$ i = 1, 2, ... p of a 10^{th} order all pole filter. The input speech samples are then inverse filtered to obtain an excitation sequence.

Processes III, IV and V are entered if the frame is declared as voiced.

Process III

In this process, the centre P^n samples of the n^{th} frame excitation sequence are employed to form one pitch cycle or a "prototype waveform". The prototype waveform is then DFT transformed to yield a vector of spectral magnitudes $\{MG^n_j\}$ j = 1, 2, ..., P^n / 2.

Process IV

In the 1.5kbit / sec coding realisation of the MPS-PIC model, the $\{MG^n_j\}$ spectral samples of process III are spectrally weighted. This weighting is determined by the shape of the $\{a_i^n\}$ LPC filter characteristic. The Euclidean norm, Si^n, of the weighted vector of $\{MG_j^n\}$ samples is then calculated. This is used during decoding to ensure that synthesised frames have the same energy as original input frames. Thus the synthesis process at the decoder uses SI^n and the LPC envelope information to represent the prototype magnitude spectrum $\{MG_j^n\}$ information.

Process V

The $\{MG_j^n\}$ samples represent the excitation magnitude spectrum sampled at pitch harmonic frequencies. In this process, each of these harmonics is declared either "fully voiced" or "mixed voiced" [3].

Using these "mixed voicing" classifications, a cut-off frequency f_c^n is calculated. Below this frequency, all harmonics are assumed to be "fully voiced", whereas harmonics located above f_c^n are assumed "mixed voiced". In general, f_c^n is determined to be the frequency which minimises the number of false classifications caused by this assumption. Note that in the mixed excitation voicing implementation of the 1.5kbits/sec system, f_c^n is allowed to take only one of four values.

Process VI

This process is carried out in the case of the frame being declared as unvoiced. Here, the energy E^n of the LPC residual signal defined from the middle of the $(n-1)^{th}$ frame to the middle of the n^{th} frame is calculated.

Following the above procedures, the n^{th} frame parameters which must be transmitted to the decoder are (1) The LPC filter coefficients $\{a_i^n\}$, (2) The pitch value P^n, (3) The voiced / unvoiced classification V^n, (4) For voiced frames, i) a "single value" spectral magnitude representation SI^n and ii) the "mixed excitation flags" cut-off frequency f_c^n and (5) For unvoiced frames, the energy E^n of the LPC residual signal.

Figure 2 depicts the synthesis process employed at the decoder. Adjacent sets of excitation parameters are used to generate an excitation sequence over the interval defined from the middle of the $(n-1)^{th}$ frame to the middle of the n^{th} frame. The sets of filter coefficients $\{a_i^{n-1}\}$ and $\{a_i^n\}$ of the $(n-1)^{th}$ and n^{th} frames are then used to synthesise speech twice over the above interval. This is followed by an overlap / add process which ensures the smooth evolution of the signal's spectral envelope information. A postfilter, which consists of a highpass element and formant enhancement is then applied on the synthesised speech, to produce the final speech output.

Notice that in the case of voiced speech, the excitation sequence consists of two components. The first of these is obtained from a bank of pitch harmonic oscillators. The amplitude and frequency of these oscillators varies and at the centre of each frame is determined from the recovered $\{MG_j^n\}$ parameters (calculated using SI^n, $\{a^n_i\}$ and P^n). In between these points, amplitudes and frequencies are calculated every sampling instant using interpolation.

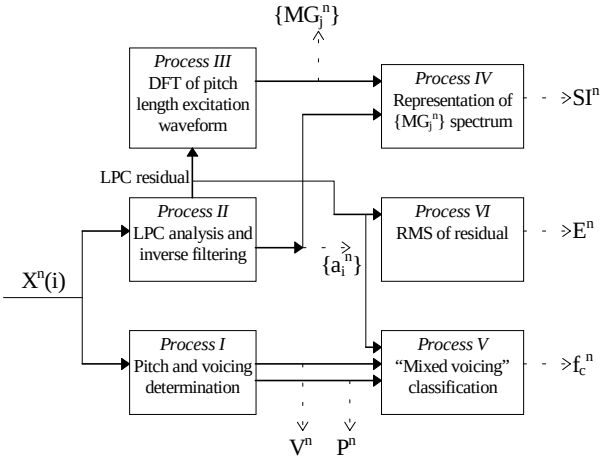

Figure 1 MPS-PIC encoding process

Thus the instantaneous phase of these "harmonic oscillators" is obtained via a polynomial and evolves smoothly. The second component is a set of "random oscillators". These are "placed" in frequency with a spacing of 50Hz around each MG_j^n harmonic which has been declared as "mixed voiced" (i.e. whose frequency is larger than f_c^n). Their magnitude is related to that of the parent harmonic. The phase of these oscillators is randomised every pitch interval. Furthermore, the excitation sequence employed for unvoiced frames is generated using random Gaussian noise whose power level is adjusted to the received power level E_n of the original LPC residual signal

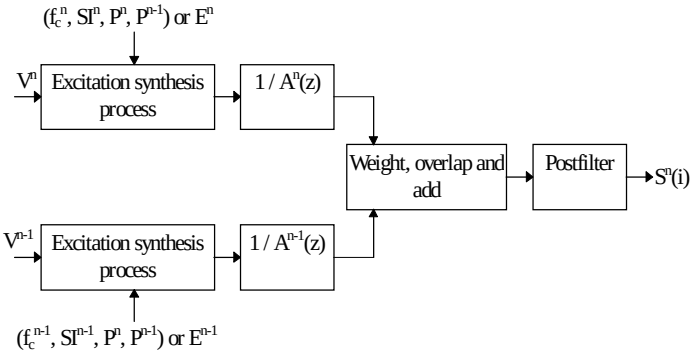

Figure 2 MPS-PIC decoding process

3. MANCHESTER SEGMENTAL QUANTISATION (MSQ)

The Manchester Segmental Quantisation model is shown diagramaticaly in figure 3. This scheme buffers sets of MPS-PIC parameters obtained from M 20msecs input frames, typically M = 10. A novel technique is then used to split such 200ms intervals into a number of variable length segments. Some of these segments are "quantised" according to the methods described in the following section and their MPS-PIC excitation parameters are transmitted. The remaining segments

(usually one), and any frames at the end of the interval which do not belong to any segment are retained and fed into the next segmentation interval. The proposed segmentation minimises, as far as is possible, the distortion produced through quantising the LSP vectors of the resulting segments whilst keeping the segment rate near to an expected phoneme rate. The aim is to break up the input speech into "sub - sounds" or "events" which can be subsequently taken as entities for quantisation purposes. Note that in this implementation, segments consist of all voiced frames or all unvoiced frames and may not contain a mixture of the two.

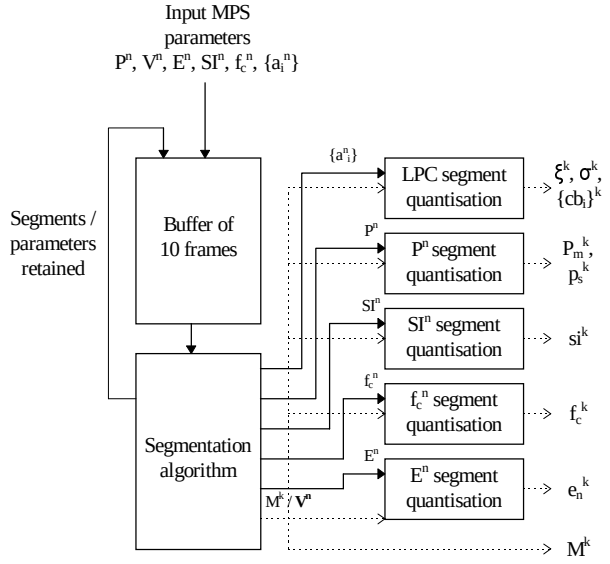

Figure 3 Segmental Quantisation model

4. MSQ TECHNIQUES

The evolutions within a segment of the $\{a_i^n\}$, P^n, E^n, SI^n and f_c^n parameters of the MPS-PIC model are quantised separately. Thus the quantisation of the parameters of the k^{th} segment of length M^k ($1 \le M^k \le L_{max}$) are described below:

LPC spectra

The $\{a_i^n\}$ i = 1, 2, … p LPC coefficients (p=10) obtained from each of the M^k input frames are converted to LSP's i.e. $\{lsp_i^n\}$. This results in p LSP "tracks" of length M^k. Each segment is "classified" using a "classification codebook". This procedure labels the segment as belonging to a specific area in the "spectral evolution space". The system then considers the evolution of separate LSP tracks with time. A set of p VQ codebooks are selected (i.e. one for each LSP track) according to the segment classification ξ^k. The length of each vector in these codebooks is fixed to L_{max}, the maximum allowable length of a segment. The quantisation of the LSP tracks then takes place as follows:

$$\lambda = \min_\sigma \left\{ \sum_{i=1}^{p} \min_j \left\{ \sum_{m=1}^{M^k} w_{m,i} (lsp_{m,i} - cb^{\xi,j}{}_{m+\sigma,i})^2 \right\} \right\} \quad (1)$$

$$\sigma = \arg^{-1} \left[\min_\sigma \left\{ \sum_{i=1}^{p} \min_j \left\{ \sum_{m=1}^{M^k} w_{m,i} (lsp_{m,i} - cb^{\xi,j}{}_{m+\sigma,i})^2 \right\} \right\} \right] \quad (2)$$

$$Cb^i = \arg^{-1} \left\{ \sum_{m=1}^{M_k} w_{m,i} (lsp_{m,i} - cb^{\xi,j}{}_{m+\sigma,i})^2 \right\} \quad (3)$$

σ = 1, 2, …, L_{max} - M^k

where $lsp_{m,i}$ (m = 1, 2, .., M^k. i = 1, 2, ... p) are the elements of the matrix of LSP vectors of the segment, $cb^{\xi,j}_{m,i}$ (j = 1, 2, …, CBS^ξ_i) are the individual elements of the i^{th} codebook for classification ξ, CBS^ξ_i is the codebook size and $w_{m,i}$ is a weighting function.

This procedure produces a quantisation distortion λ^k for the segment, a "codebook displacement" σ^k and a set of p codebook indices $\{Cb_i^k\}$. ξ^k, σ^k and $\{Cb_i^k\}$ are then transmitted to the decoder. The sets of allowable classifications for voiced and unvoiced segments are separate and thus ξ^k also implicitly represents the voicing status of the k^{th} segment.

LSP vectors are recovered using:

$$\overbrace{lsp_{m,i}} = cb^{\xi, Cb_i}{}_{m+\sigma,i} \quad (4)$$

Notice that the displacement σ can take any value, as long as $\sigma + M^k$ is not greater than the length L_{max} of the codebook vectors and that all of the p LSP tracks share the same displacement σ but have different $\{ Cb_i^k \}$ quantisation indices.

Also note that during the codebook search process described in equations (1) to (3), an additional constraint on the choices of σ^k and $\{Cb_i^k\}$ is applied to ensure the correct ordering of the reconstructed LSP vectors and thus guarantee the stability of the synthesis filter.

Pitch

For the k^{th} segment, the pitch information consists of a vector $\mathbf{P}^k$ of length M^k. The mean of this vector, $P_{mean}{}^k$ is calculated and subtracted from $\mathbf{P}^k$ to form a zero mean vector of pitch values $\mathbf{P_{mr}}^k$. $P_{mean}{}^k$ is also quantised differentially using a single tap predictor, to give a quantisation index $p_m{}^k$. $\mathbf{P_{mr}}^k$ is vector quantised using a "shape" codebook whose vectors are of length M^k giving an index $p_s{}^k$. Notice that a "family" of "shape" codebooks is employed with one codebook for each possible M^k.

Unvoiced frame energy $\boldsymbol{E}^k$ and single value amplitude representation $\boldsymbol{SI}^k$

Again, for each segment, these parameters consist of a single vector of length M^k, which is quantised in the same manner as $\mathbf{P}^k$. Thus in unvoiced frames, $\mathbf{E}^k$ is represented by a mean index $e_m{}^k$ and a shape index $e_s{}^k$ whereas in voiced frames $\mathbf{SI}^k$ is represented by $si_m{}^k$ and $si_s{}^k$.

"Mixed voicing" transition frequency $\boldsymbol{f_c}^k$

Given a transition frequencies vector of length M_k, no mean is extracted from the vector whose "shape" is simply transmitted in the form of an index $f_s{}^k$. The shape quantisation scheme is the same as that applied to $\mathbf{P_{mr}}^k$.

Notice that in addition to the above quantisation indices, the segment length M^k is also transmitted.

5. RESULTS AND DISCUSSION

The segmentation and quantisation schemes described in sections 3 and 4 contain a large number of interdependent parameters which affect the reconstructed speech quality and bitrate. In this paper we consider the performance of MPS - PIC segmentally quantised schemes operating at 1.0kbits/sec, 900bits/sec, 800bits/sec and 700bits/sec. The parameters of these schemes were experimentally optimised via informal subjective tests. In addition, a MPS-PIC 1.5kbits/sec fixed bit rate system employing frame by frame quantisations was tested and compared with the above average bit rate schemes. The 1.5kbits/sec system incorporates mechanisms which exploit interframe redundancy and an LSP split matrix quantisation approach that attempts to preserve the smooth evolution of the LPC filter parameters [10]. No attempt is made to ensure the smoothness of the evolutions of the other Codec parameters. Notice that the true instantaneous bit rate of these Segmental PIC schemes is variable and depends upon the speaker and the dynamics of the speech. In the context of these results, however the term 'bitrate' refers to the average rate measured over 30 minutes of input speech material comprising of various speakers and messages.

These five schemes were evaluated using informal subjective tests. For each test, the participants listened to a reference sentence, generated from the MPS-PIC model with no parameter quantisation, followed by the same sentence MPS-PIC processed while using the quantisation scheme under investigation. Subjects were asked to give for each test sentence a score which indicates it's similarity to the reference according to the scale of table 1.

1. The difference between the files is significant and disturbing. The second file sounds "unnatural"
2. The difference between the files is perceivable and significant, but the second file still sounds "natural"
3. The difference between the files is clearly perceivable, but only minor
4. The difference between the files is just about perceivable
5. There is no perceivable difference between the files

Table 1 Subjective testing scoring scale

Several sentences, each uttered by a different speaker were used in these tests which involved 8 subjects.

Mean subjective score results are shown in figure 4. Notice that the 1.5kbits/sec scheme is not "transparent" when compared to the unquantised model. In addition, some subjects commented that although the output of the 1.0kbps and 900bps schemes differed slightly from that of the unquantised model, they felt that the quantised model was in fact preferable to the unquantised reference. These results and comments reflect the difficulty in attempting to subjectively judge the difference between utterances which sound very similar. However, these informal subjective tests also indicate that the proposed generic segmental quantisation approach can almost halve the bit rate of a Prototype Interpolation Coding systems (e.g. MPS-PIC) while preserving the overall recovered speech quality.

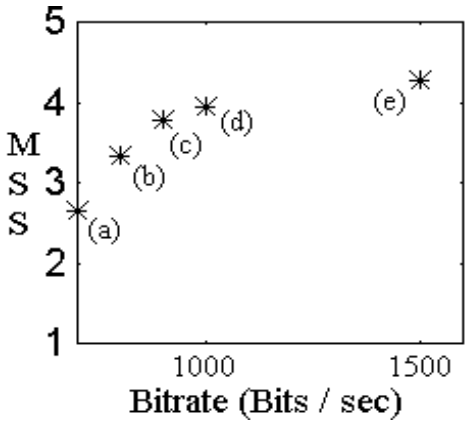

Figure 4 Mean Subjective Scores (MSS) for the test schemes at (a) 700 BPS (b) 800 BPS (c) 900 BPS (d) 1.0kbps and (e) 1.5kbps

6. REFERENCES

1. Kleijn W. B. , Haagen J., *"Waveform interpolation for coding and synthesis"* Speech coding and synthesis 1995 Elsevier Science Chapter 5 pp 175 - 207
2. Xydeas C. S., *"Speech synthesis system"* Patent No. GB97/01831
3. Xydeas C. S. , Papanastasiou C. , *"Efficient mixed excitation models in LPC based prototype interpolation speech coders"* Proc IEEE - ICASSP 97 pp 1555- 1558
4. Xydeas C. S. , Cao B. *"Source driven Variable bit Rate Prototype Interpolation Coding"*, Proc IEEE - ICASSP 96 pp 220 - 223
5. McAulay R. J. , Quatierri T. F. , *"Speech analysis / synthesis based on a sinusoidal representation"* IEEE ASSP Vol 34 No 4 August 1986 pp 744 - 754
6. Griffin D. W., Lim J. S. , *"Multiband excitation vocoder"* IEEE ASSP Vol 36 No 8 August 1998 pp 1223 - 1235
7. Tokuda et al. *"A very low bit rate speech coder using HMM based speech recognition / synthesis techniques"* Proc IEEE ICASSP 98 Vol 2 pp 609 - 612
8. Cernocky . J, Baudion G. , Chollet G. , *"Segmental vocoder - going beyond the phonetic approach"* Proc IEEE ICASSP 98 Vol 2 pp 605 - 608
9. Honda M. , Shiraki Y. , *"Very low bit rate speech coding"* Advances in speech signal processing Editors S. Furui, M. M. Sondhi Markel Dekker 1991 Chapter 6
10. Xydeas C. S. , Papanstasiou C. *"Efficient coding of LPC parameters using split matrix quantisation"* Proc IEEE - ICASSP 95 pp 740 - 743
11. Xydeas C. S. , Chapman T. M. *"Multi codebook Vector Quantization of LPC parameters"* Proc IEEE ICASSP98, Vol 1 , pp 61 -64