Proceedings of the Institute of Acoustics

A LOW COMPLEXITY, VARIABLE RATE SPEECH CODER FOR DIGITAL TELEPHONE ANSWERING MACHINES

C. I. Parris (1), D Y K Wong (1) and G Faure (2)

(1) Ensigma Ltd., Chepstow, Gwent

(2) Telecom Paris

1. INTRODUCTION

Two important characteristics separate speech storage algorithms from those for speech transmission. The first is that the encoding bit rate can vary so as to maintain a uniform reproductive quality, and the second is there is no limit on encoding latency subject to adequate buffering. This allows wider optimization windows.

These properties can be exploited to increase coding efficiency relative to telecommunications coders.

In this paper a variable rate speech coding algorithm is presented which performs a phonetic segmentation of the speech message. Four phonetic classifications are used, namely silence, unvoiced, mixed voiced and strongly voiced. Each segment is efficiently encoded using code excited linear predictive (CELP) techniques. The average bit rate for each segment class is constrained to reflect their respective perceptual importance.

The perceptually most important phonetic segments are strongly voiced. In such segments the LPC residual in each pitch period consists of a large major pulse surrounded by a number of smaller pulses. Recent work by Atal [1] and Shoham [2] indicates that CELP coders typically fail to model the periodicity of voiced speech since there are conflicting requirements for the excitation codebook to model the dominant glottal impulse, particularly at the onset of the voice segment, and to model the large number of time varying secondary pulses which occur around the main pulse. More recent work has attempted to address this problem by designing hybrid excitation codebooks which incorporate an area populated by sparse large amplitude pulse like vectors and an area populated with more random noise like vectors. However, Granzow [3] states that the selection of pulse vectors from a codebook, based on individually encoding relatively short speech blocks, usually does not result in an excitation with a smoothly evolving pulse interval. For obtaining such a consistent periodic excitation, large optimization frames are required.

In this paper we present an enhancement to CELP which allows the dominant single pulse excitation train to evolve alongside a normal CELP excitation for regions of voiced speech. This is achieved by centring a prototype pitch excitation waveform around each pitch marker. The pitch markers are determined over a wide observation window and are used to segment the excitation into sub-frames centred on each pitch marker. This allows a quasi-periodic excitation to be constructed by successively adapting a prototype pitch excitation waveform, thereby improving periodicity.

2. PHONETIC SEGMENTATION

Initially analysis proceeds as in a normal low bit rate CELP coder. We use an 8 kHz sample rate and a 32 ms frame size with four 8 ms sub-frames. The short-term Linear Predictive (LP) analysis is performed once per 32 ms frame by open loop, 10th order autocorrelation analysis using a 32 ms Hamming window, no pre-emphasis and 15 Hz bandwidth expansion. The analysis window is centred at the end of the last sub-frame. The linear predictor is coded using scalar quantization of line spectral pairs (LSP). A 34-bit absolute encoding is performed but the resultant LSPs may be differentially encoded provided that the differential encoding results in the same LSP set. A 24-bit two dimensional differential encoding scheme is used, similar to the technique presented in [3]. Lower bit rates are possible using vector quantization (VQ) of the LSP set, however this significantly increases computational and memory requirements. The encoding scheme achieves less than 1.5 dB^2 of spectral distortion.

Open loop pitch prediction is then performed on the perceptually weighted short-term speech residual. Let $C_o(k)$ be the correlation corresponding to an integer lag k in the open loop sense.

$$C_o(k) = \sum_{n=0}^{N-1} w(n)w(n-k), \qquad \text{for } k = L_{min}, L_{max}$$

and define $G_o(k)$ as :

$$G_o(k) = \sum_{N=0}^{N-1} w^2(n-k), \qquad \text{for } k = L_{min}, L_{max}$$

where $w(n)$ is the spectrally weighted input speech, N is the number of samples in a sub-frame and L_{min} and L_{max} specify the range of integer lags. J, the lag which maximizes the prediction gain of a first order integer lag pitch predictor over the weighted speech for the sub-frame, can be found by setting J to the value of k which maximizes the normalized correlation :

$$\frac{C_o(k)}{\sqrt{G_o(k)}} \qquad k = L_{min}, L_{max}$$

Once J, the best integer lag, has been obtained, sub-multiples of J are evaluated to see if they are normalized correlation peaks. If the sub-multiple peak exceeds a specified percentage of the maximum it is excepted as the true integer lag. This process is repeated to obtain the second best integer lag. The pitch estimates for consecutive sub-frames are examined for consistency allowing for a 10% variation in pitch from sub-frame to sub-frame. The open loop pitch prediction gain P_i in addition to the rms signal level $\sqrt{S_i}$ are used to assign a tentative classification for the ith sub-frame. P_i and S_i are given by

$$S_i = \sum_{n=0}^{N-1} s^2(n), \qquad P_i = 10\log_{10}\left[\frac{S_i}{S_i - \frac{C_o(J)}{G_o(J)}}\right]$$

where s(n) is the speech sub-frame under consideration. The classification algorithm is as follows - if (S_i <SILENCE_THRESHOLD) classification = SILENCE, else if ($P_i < 3dB$) classification = UNVOICED, else classification = VOICED.

The silence threshold can be determined using the background line level of the telephone, directly from a segment of speech or a combination of both. The threshold is improved as more speech is processed.

Consecutive sub-frames which have the same classification are considered for chaining into a sub-frame list. Consecutive silence or unvoiced sub-frames may be chained directly. Voiced sub-frames are only chained provided the pitch period is consistent from sub-frame to sub-frame. The minimum length of a valid chain is three sub-frames and the maximum chain length is unrestricted - this does imply excess buffering since the chain can be processed as it forms.

After chaining, the remaining unchained sub-frames are grouped together, chained and given the classification of either unvoiced or mixed voiced. Unvoiced classification is applied to only speech segments consisting of silence and unvoiced sub-frames. Mixed voiced classification is applied to segments which contain some voiced sub-frames.

Silence, unvoiced and mixed voiced lists are processed directly, with each sub-frame undergoing an analysis-by-synthesis process. Voiced lists are first processed to determine the location of the pitch markers within the speech segment. The pitch marker locations are then used to derive a new set of sub-frame boundaries for the voiced sub-frame list. The boundaries are located at the mid-points between successive pitch markers. Figure 1 illustrates the LPC analysis and front end processing for the speech storage algorithm. In the cases of silence, unvoiced and mixed voiced speech segments, the sub-frame boundaries are unchanged compared to the original 8 ms sub-frames. The output of this processing stage is the perceptually weighted short-term speech residual vector W, the dimension of this vector being the sub-frame size.

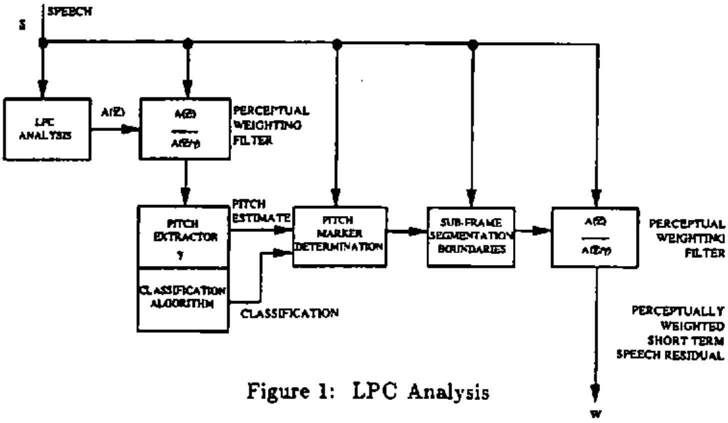
SPEECH
s
LPC ANALYSIS
A(Z)
A(Z)
A(Z/γ)
PERCEPTUAL WEIGHTING FILTER
PITCH EXTRACTOR
γ
CLASSIFICATION ALGORITHM
PITCH ESTIMATE
CLASSIFICATION
PITCH MARKER DETERMINATION
SUB-FRAME SEGMENTATION BOUNDARIES
A(Z)
A(Z/γ)
PERCEPTUAL WEIGHTING FILTER
PERCEPTUALLY WEIGHTED SHORT TERM SPEECH RESIDUAL
W

Figure 1: LPC Analysis

3. PITCH MARKER DETERMINATION

The sub-frame list is segmented into observation windows of four sub-frames (32 ms). These windows overlap by one sub-frame where possible, with the fourth sub-frame of one window corresponding to the first window of the next. The pitch markers are successively determined for each window, but only those markers that fall into the first three sub-frames of the window are retained.

The pitch marker corresponds to the major excitation pulse. It occurs at the instant of glottal closure. In this paper we assume that this excitation can be approximated by a single impulse located at the pitch marker location. Previous work [1] has shown that good speech quality can be achieved if periodic speech is synthesized by exciting an all-pole LPC filter with only a single such excitation per pitch period. By combining this excitation in addition to a usual CELP excitation high quality speech is obtained. To locate each pitch marker we use a modified version of the excitation point finding algorithm given in [4].

The first step is to search the speech in the observation window for the global maximum. The hypothesis is that this maximum results from a single impulse at the pitch marker location. To locate this pitch marker we compute the maximum signal to noise ratio that can be obtained by exciting the LPC filter with a single optimally scaled impulse located anywhere in a window relative to the global maximum. Denoting the first n samples of the impulse response by $\boldsymbol{h}_n$ we define an error vector $\boldsymbol{e}_{m,n}$ as the difference between the speech vector $\boldsymbol{s}_{m,n}$ and the optimally scaled impulse response.

$$\boldsymbol{e}_{m,n} = \boldsymbol{s}_{m,n} - \alpha \boldsymbol{h}_n$$

where

$$\boldsymbol{s}_{m,n} = [s(m), s(m+1), ...s(m+n)]$$

and

$$\alpha = \frac{\boldsymbol{s}^T_{m,n}\boldsymbol{h}_n}{\boldsymbol{h}^T_n \boldsymbol{h}_n}$$

$\boldsymbol{s}_{m,n}$ represents the n samples of speech from sample m in the observation window. If the global maximum occurs for $m = M_o$ then $\boldsymbol{e}_{m,n}$ is evaluated for m in the range $[M_o - 0.1P, M_o + 0.1P]$ where P is the pitch estimate.

The power in the error signal is given by

$$\boldsymbol{e}^T_{m,n}\boldsymbol{e} = \boldsymbol{s}^T_{m,n}\boldsymbol{s}_{m,n} - \frac{\left(\boldsymbol{s}^T_{m,n}\boldsymbol{h}_n\right)^2}{\boldsymbol{h}^T_n \boldsymbol{h}_n}$$

The pitch marker location occurs at position $m = p_o$ which maximizes the signal to noise ratio $S_{m,n}$.

Subsequent pitch markers are located in a similar manner by considering other local maxima in the speech signal in regions which are approximately an integer pitch period away from the global maximum.

According to [4] the value n is chosen to be one third of the pitch period estimate.

Once all the pitch markers have been located a consistency measure is calculated from the signal to noise values, the pitch marker amplitudes and inter-marker intervals. This measure is similar to the cost function derived in [3] and is empirically derived. Any inconsistent pitch marker sets are assumed not to fit our model and are processed as mixed voiced speech.

4. EXCITATION CODING

The rationale behind phonetic segmentation is to apply the most appropriate excitation encoding scheme to each segment. Figure 2 illustrates the scheme.

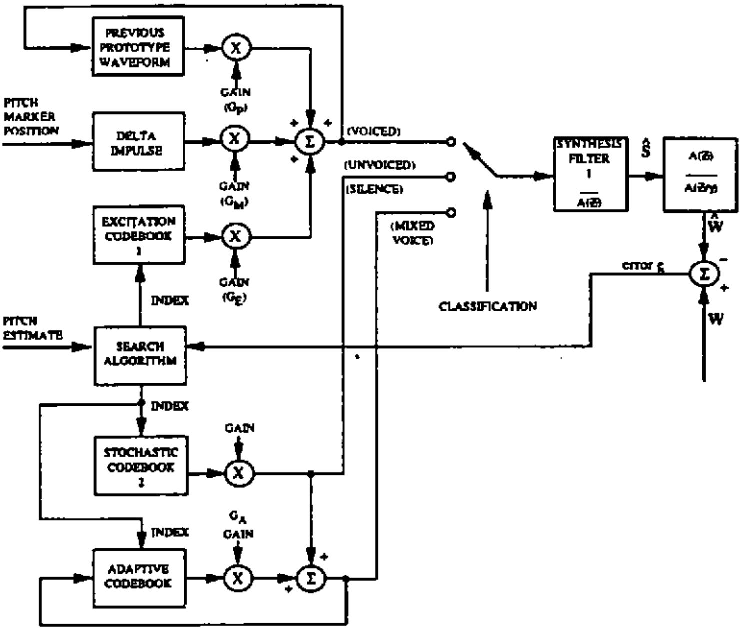

Figure 2: Excitation Coding

In this context silence corresponds to very low energy segments. The LPC excitation for these regions is a gain scaled stochastic vector. The vector is selected at random from a stochastic codebook, hence no codebook index is required. The gain is encoded in 5 bits and is selected to achieve the same energy in the synthetic sub-frame as occurs in the original. The sub-frame size is 8 ms and the gain is updated at the sub-frame rate. This results in a bit rate of 625 kb/s for silence segments. Any silence segments which occur at the start or end of the message may be ignored.

Since unvoiced segments have a noise-like waveform with fairly weak near-sample correlation the long-term predictor is omitted. The LPC excitation is a gain scaled stochastic codebook vector. Excitation vectors with dimension 64 are selected from a stochastic codebook containing 512 vectors. We experimented with bit allocations for unvoiced sections to find the minimum bit rate at which a further increase in rate would cause a barely perceptual improvement. This approach was also taken in designing most other components of the algorithm. The result is an overall bit rate of 1.625 kb/s to encode the excitation.

Mixed voiced segments are characterized by either a weak or rapidly varying pitch content. Whenever we cannot perform adequate pitch prediction the resulting residual signal may have many large pulses. It is extremely difficult to correctly render these pulses as well as the periodicity with a codebook constructed of noise-like code vectors. Thus a combination of adaptive codebook and large excitation codebook is used. Eight bits are used to encode the adaptive codebook index and five bits for gain. The codebook structures are identical to those given in the DoD 1016 Standard, with the modification to 8 ms sub-frames. The excitation codebook consists of a single pulse section comprising 64 elements and a stochastic codebook comprising 448 elements. Again five bits are used to encode gain. This results in a bit rate of 3.375 kb/s for the LPC excitation in mixed voiced segments.

The LPC excitation in periodic voiced speech segments is a slowly varying prototype pitch excitation waveform. This waveform is of dimension equal to the instantaneous sub-frame size and is centred on the pitch marker location for the current pitch period. For the *i*th sub-frame three vectors are combined to derive the excitation vector $\boldsymbol{V}_i$

$$\boldsymbol{V}_i = \boldsymbol{V}_{i-1} * G_P + \boldsymbol{V}_E * G_E + \boldsymbol{V}_M * G_M$$

where

$\boldsymbol{V}_{i-1}$	=	Previous prototype waveform
G_P	=	Gain value
$\boldsymbol{V}_E$	=	Excitation codebook vector
G_E	=	Gain value
$\boldsymbol{V}_M$	=	Single delta impulse vector at pitch marker location
G_M	=	Gain value

Notice that periodicity is modelled by consideration of the pitch marker locations alone, which ensures a smooth pitch contour for the dominant excitation. The pitch marker locations are differentially encoded using between 3 and 5 bits depending on pitch. The previous prototype waveform is aligned such that its pitch marker location overlays the current sub-frame's marker location. In addition any samples of $\boldsymbol{V}_{i-1}$ not within the current sub-frame are ignored or zero padded. The gain term G_P is calculated using

$$G_P = \frac{V_{i-1} W^T}{V_{i-1} V_{i-1}^T}$$

and quantized to 5 bits. The excitation vector V_E is obtained in the usual way from a codebook consisting of single pulse excitation and stochastic vectors, the vectors are modified to contain a zero at the pitch marker location. Nine bits are used to encode the V_E codebook index and 5 bits for gain. Since the pitch marker location is known the vector V_M is not encoded. The associated gain G_M is encoded using 5 bits. G_M is obtained from

$$G_M = \frac{V_M (W - G_P V_{i-1} - G_E V_E)^T}{G_P^2 V_{i-1} V_{i-1}^T}$$

Since the sub-frame rate is determined by the pitch a potentially large variation in bit rate will occur if all the above parameters are updated each sub-frame. This would also result in superior performance for high pitched speakers. To counter this the update rate for the parameters is related to the sub-frame size. Linear interpolation of all parameters between updates is used. The update rate is selected to restrict the bit rate to a maximum of 4 kb/s for encoding of the LPC excitation.

5. COMPLEXITY

The complexity of this coder is dominated by the codebook searches. The adaptive codebook search is restricted to the six closest lag values around the open loop pitch predictor lag, which significantly reduces computation. By using well structured excitation codebooks, such as VSELP [5] or ACELP [6] coding complexity is approximately 10 MIPS.

6. CONCLUSIONS

A variable rate speech coding algorithm for Digital Telephone Answering Machines and other speech storage applications has been presented. Phonetic segmentation is used to optimize bit rate. A novel pitch synchronous excitation encoding algorithm for periodic voiced segments was described, and this technique addresses some of the limitations of low-rate CELP type speech coders. Very good speech quality is achieved at average bit rates around 5 kb/s.

7. REFERENCES

[1] B S Atal and B E Caspers, 'Beyond Multipulse and CELP Towards High Quality Speech at 4 kb/s', Advances in Speech Coding, Kluwer Academic Publishers, 1991, pp. 191-201.

[2] Y Shoham, 'Constrained-Stochastic Excitation Coding of Speech at 4.8 kb/s', Advances in Speech Coding, Kluwer Academic Publishers, 1991, pp. pp. 339-348.

[3] W Granzow, B S Atal, K K Paliwal and J Schroeter, 'Speech Coding at 4 kb/s and Lower Using Single-Pulse and Stochastic Models of LPC Excitation', ICASSP '91, Toronto, pp. 217-220.

[4] K Y Lo, B M G Cheetham, W T K Wong and I Boyd, 'A Pitch Synchronous Scheme for Very Low Bit Rate Speech Coding', IEE Colloquium on Speech Coding Techniques and Applications, April '92, pp. 3/1-3/5.

[5] I A Gerson, 'Techniques for Improving the Performance of CELP Type Speech Coders', IEEE Journal on Selected Areas in Communications, June '92, Volume 10, No. 5, pp. 858-865.

[6] C Laflamme, J P Adoul, R Salami, S Morissette and P Mabilleau, '16 kb/s Wideband Speech Coding Technique Based on Algebraic CELP', ICASSP '91, Toronto, pp. 13-16.