SOURCE CONTROLLED VARIABLE BIT-RATE SPEECH CODER BASED ON WAVEFORM INTERPOLATION

F. Plante(*), B.M.G. Cheetham(*), D. Marston (o), P.A. Barrett(+)

(*) Dept. Electrical Engineering & Electronics, Liverpool University, Liverpool L69 3BX, UK

(o) Ensigma Ltd, Turing House, Station Rd, Chepstow, NP6 5PB, UK
(+) BT Laboratories, Martlesham Heath, Ipswich IP5 7RE, UK

ABSTRACT

This paper describes a source controlled variable bit-rate (SC-VBR) speech coder based on the concept of prototype waveform interpolation. The coder uses a four mode classification : silence, voiced, unvoiced and transition. These modes are detected after the speech has been decomposed into slowly evolving (SEW) and rapidly evolving (REW) waveforms. A voicing activity detection (VAD), the relative level of SEW and REW and the cross-correlation coefficient between characteristic waveform segments are used to make the classification. The encoding of the SEW components is improved using a gender adaptation. In tests using conversational speech, the SC-VBR allows a compression factor of around 3.
The VBR coder was evaluated against a fixed rate 4.6kbit/s PWI coder for clean speech and noisy speech and was found to perform better for male speech and for noisy speech.

1. INTRODUCTION

One of the primary motivations for using variable bit-rate (VBR) transmission is to increase the efficiency of bandwidth utilisation. Indeed, most examples of existing VBR schemes are intended to increase the number of users that can be supported within a fixed bandwidth. This may be because the bandwidth is relatively expensive, as in the case of international trunk circuits, or because bandwidth is very scarce, which is usually the case in a mobile radio system.
Speech inherently lends itself to VBR coding because the information rate varies substantially with time and for a given level of quality and intelligibility, the minimum number of bits required to faithfully encode a segment of speech will vary from frame to frame. The basic elements of conversational speech can be broadly classified into four categories: voiced speech, unvoiced speech, transitions and silence [1]. The term 'silence' is rather misleading and should be taken to mean periods of talker inactivity of which true silence is a special case. This distinction is particularly important for a mobile radio system where users are likely to be in a noisy acoustic environment.

In theory, a different coding technique can be used for each class of speech. Methods by which the mode can be selected for a given segment of speech may be described as open loop or closed loop [2]. In the open-loop approach, a classification decision concerning the speech frame is made prior to the encoding process completed in a single pass. This places great emphasis on the reliability of the classification decision. Incorrect classification decisions degrade the transmission quality.

With closed-loop control a speech frame is encoded and decoded using each of the coding methods available. The quality of each decoded speech frame is evaluated using an objective speech quality measure and compared to a threshold. The encoded speech is transmitted using the method which exceeds the required quality threshold. The closed-loop technique is inherently more robust than the open-loop technique because it takes into account the effects of encoding and decoding the particular speech frame. However, the closed-loop approach has two serious drawbacks. Firstly, the need to encode and decode speech frames at each coder mode substantially increases the complexity of the coder. Secondly, it is difficult to evaluate the quality of the candidate speech frames at each mode in a way which reliably predicts the preference of a human listener.

The VBR coding scheme proposed in this paper has the advantage of an open loop method in that only one encoding process is applied to each speech frame. However the classification is made after the parametrisation of the speech. This is possible because of the principle of prototype waveform interpolation (PWI) which decomposes the speech into periodic and random components. The choice of bit-rate and parameters are made at the quantisation stage (figure 1).

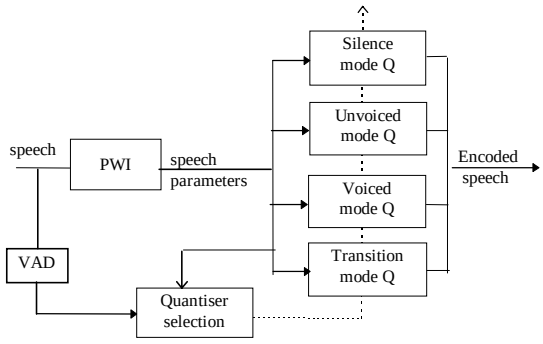

Figure 1 : Scheme of the source control variable bit rate PWI encoder

The variable bit-rate coder is based on a fixed rate 4.6 kbit/s version of PWI.

2. PWI

The prototype waveform interpolation approach aims to cater for the non-stationarity of speech and to take advantage of the way it is perceived by decomposing the changes that occur to pitch-cycle length residual segments (referred to as characteristic waveforms) into slowly evolving (SEW) and rapidly evolving (REW) waveform components [5]. The SEW components arise from the quasi-periodic content of the speech residual. The REW components arise from degrees of randomness embedded in all forms of speech but particularly in unvoiced speech. For voiced speech segments, the SEW components will tend to be dominant whereas for unvoiced speech, the REW components will be more important.

The coder described is based on a 10 ms frame size, which necessitates a 40ms look-ahead, mainly due to the pitch detection. The linear prediction coefficients are estimated using pitch synchronous Burg LP analysis and converted to line spectrum frequencies (LSF). The LSFs are then quantised using the G.729 '18-bit' differential multi-stage split vector quantiser [3]. A residual signal is obtained using an FIR analysis filter with the quantised LSF coefficients. Then a residual segment of pitch-period length is extracted every 2.5ms, time-normalised and aligned to form a two-dimensional surface. To extract the SEW component the surface is low-pass filtered along the time-domain and down-sampled at intervals of 10ms. The SEW component is DFT transformed to obtain magnitude and phase spectra.

To take into account properties of the human ear, the SEW and REW spectra are smoothed into critical frequency bands [8]. It has been shown that more efficient vector quantisation could be achieved using a gender specific scheme [9]. For the SEW, the number of critical bands depends of the gender classification (16 for female, 32 for male). The smoothed SEW magnitude spectrum is then vector quantised using a multistage vector quantisation with 7 and 4 bits. The quantisation is adapted to the gender class. For the REW, 16 bands are used. The vector is then vector quantised using a 7 bit codebook.

At the decoder, the SEW magnitude spectrum is reconstructed using cubic spline interpolation, and normalised to the pitch length. A second order all-pass filter [6] is used to approximate the phase spectrum of the SEW. The all-pass filter artificially reproduces some of the characteristics of the glottal excitation. The coefficients of the all-pass filter are pitch dependent.

The REW magnitude spectrum is recovered from the cepstral coefficients and interpolated to obtain a REW magnitude every 2.5ms. At each sub-update point, a random phase spectrum is defined with each phase uniformly distributed between π and $-\pi$. The complex spectrum of the SEW and REW are added, and a residual signal recovered using quadratic interpolation [4].

Then the vocal tract resonances are adding using an IIR synthesis filter with the LSF coefficients.

3. SOURCE CONTROLLED VBR

3.1 Voiced Activity Detection

A voice activity detector (VAD) is used to distinguish speech from background noise. The VAD used is based on the GSM voice activity detector as adapted to the G.729 speech coder standard [7]. It gives a detection decision every 10 ms, and has been shown to give accurate detection of speech in noisy environments [7]. It consists of two separate detection units, the primary and secondary VAD's. Using an inverse filter, the primary VAD attempts to filter out any background noise which is assumed to be stationary and non-periodic, and the energy of the filtered signal is compared with an adaptive threshold to decide whether speech is present. This provides robust detection in the presence of high background noise levels. The secondary VAD decides when to update the coefficients of the inverse filter and adapt the primary VAD threshold. This occurs when the input is classified as stationary and non-periodic.

When the VAD indicates that speech is not present, the silence mode is selected. A random signal is generated at the receiver to provide "comfort noise" between speech bursts. To ensure good continuity in the reconstructed signal, the energy and frequency spectrum of the background noise are matched. This is done by transmitting the REW component.

3.2 Speech classification

As mentioned previously, the changes in the characteristic waveform will depend of the nature of the speech. For unvoiced speech, there will be a higher REW component whereas for voiced speech, there will be a higher SEW component. The relative level of SEW and REW components could be considered as a good indicator of the nature of the speech. The classification between the voiced, unvoiced and transition modes are based on the energy ratio between the SEW and REW and the mean normalised cross-correlation coefficient obtained during the alignment stages. These two parameters are summed to form a mode factor.

Unvoiced mode:

When the mode factor is lower than 1.5 or the energy ratio lower than 0.85 the frame is classified as unvoiced. If the previous frame was unvoiced, these thresholds are increased to 1.7 and 0.9 respectively.

During unvoiced speech, the periodic component of the signal is discarded. The pitch and SEW are discarded. Only a coarse description of the LSFs are needed, so only the first 8-bits of the G.729 quantiser are used.

Voiced mode:

When the mode factor is greater than 2.0, the frame is classified as voiced. If the previous frame was already voiced, the threshold is decreased to 1.8. The REW component is discarded. The LSF coefficients are quantised using the two codebooks of the G.729 quantiser. The pitch is scalar quantised

with 7 bits, and the SEW magnitude uses an 11-bit vector quantiser.

Transition mode:

When the speech frame is neither classified as voiced or unvoiced, the transition mode is selected. In this mode both periodic and random components are of perceptual importance. All the speech parameters are encoded and quantised using the higher bit-rate.
A transition frame is imposed when passing from a voiced frame to an unvoiced or silence frame. This will increase the average bit-rate, but removes artefacts at these transitions.

Gender Selection:

The gender is determined from pitch estimates. A running average is taken over sections of voiced speech, and the if average is above or below a certain threshold the gender is selected. If a positive decision can not be made the gender is considered to be 'either' and the speech coder compromises its performance for all speakers. When there are periods of speech inactivity, the gender identification system is reset due to possibility of a change of speaker.

Table 1 indicates the parameters encoded in each of the modes with the corresponding bit-rate. Currently, to simplify the transition between modes, the codebooks are the same for all the modes. In the future, bit efficiency could be gained by using a specific codebook for each mode. The gender mode is only sent when the source signal is classified as voiced or transition speech. The gender determines how the SEW is quantised which has different dimensions for male and female speech. The SEW codebooks are trained only from speech of the specific gender

	Mode (bits)			
Parameters	Silence	Unvoiced	Voiced	Transition
LSF	-	8	18	18
Pitch	-	-	7	7
SEW	-	-	11	11
REW	7	7	-	7
Modes	2	2	4	4
Bit rate (bit/s)	900	1700	4000	4700

Table 1 : Encoded parameters and bit-rate for the four modes of the coder.

4. RESULTS

4.1 Mode selection

Figure 2 represents the evolution mode selection for a male speaker sentence "the scheme was plotted out". The higher the mode value, the more the voiced component is present. It can be seen that the algorithm detects well the voiced part of the sentence. Most of the word "the" is classified as a transition because of the variation of the pitch-period in this short segment.

Back-ground	Silence	Unvoiced	Voiced	transition
clean	70.7%	13.6%	13.2%	2.5%

Table 2 : Percentages of each mode obtained with a two way conversational speech.

Table 2 gives the percentage of occurrence of each mode for examples of clean speech obtained by processing 90 seconds of both way conversational speech. The average bit-rate obtained is 1513 b/s, i.e. a compression factor of about 3.1.

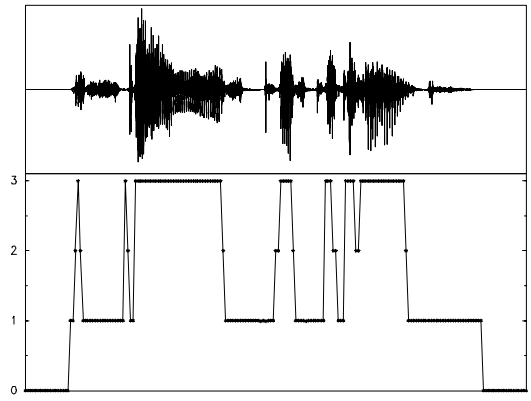

Figure 2: Speech (top) & mode classification (bottom) for a male speaker sentence: "The scheme was plotted out"

4.2 Listening tests

Listening tests have been performed to evaluate the quality of the source controlled VBR coder in comparison with a 4.6kbits/s fixed rate version of PWI.
In the tests, ten listeners were presented with 36 sample pairs. Each pair consisted of sentences obtained with the fixed rate coder and the VBR coder. The presentation of the sentence in each pair was random. The listeners were asked to make a three way choice for each sentence pair : first one better, no preference, second one better. Preference scores obtained for different types of background noise are given in Table 3. Figure 3 shows the preference scores obtained for male and female speech.

Background noise	Fixed Rate	No	SC-VBR
clean	56 %	34 %	10 %
car noise	52 %	28 %	20 %
babble noise	46 %	32 %	22 %
All	51 %	31 %	18 %

Table 3 : Preference scores for the coders in the different background noise.

Overall, the listeners had a preference for the fixed bit rate coder (50%). In a third of the cases, listeners found the quality of both coders comparable. However the results changed significantly when background noise was added. For car noise and babble noise, the preference for the SC-VBR coder increased by 10% and 12% respectively.

In figure 3 it can be seen that the results also depend on the gender of the speakers. For male speakers, the listeners decided that the two coders were comparable as often as they decided that the fixed rate version was better. The SC-VBR was preferred 23% of the time. For female speaker the listeners preferred clearly the fixed bit-rate with a percentage of 62%.

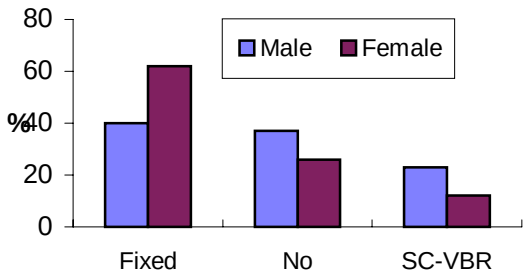

Figure 3 : Preference score of the coders for male and female speakers.

5. CONCLUSION

The principle of PWI coding allows a source controlled variable bit-rate coder to be devised without the disadvantage of the pre-classification normally required in a open loop approach, and also without the complexity of closed loop approach. The variable bit-rate scheme was found to reduce the average bit rate by a factor of about 3. Although the 4.6kbit/s fixed bit-rate version was generally preferred, the SC-VBR coder performs well and seems promising especially for male speech and for noisy speech.

6. ACKNOWLEDGEMENTS

This work is supported by EPSRC/DTI Link Personal Communications Programme. The authors wish to thank S. Watson at Liverpool University for help on the implementation of the VAD.

7. REFERENCES

1. J.L. Flanagan "Speech analysis, synthesis and perception" Second Ed. Springer-Verlag, 1972.

2. A. Gersho, E. Paksoy, "An overview of variable rate speech coding for cellular networks" IEEE ICWC92, pp.172-175, 1992.

3. ITU Recommendation C.729, "Coding of speech at 8kbit/s using conjugate structure algebraic code excited linear predictive (CS-ACELP) coding", 1995.

4. W.B. Kleijn, J. Haagen, "Transformation and decomposition of the speech signal for coding", IEEE Signal Processing letter, Vol.1, pp.136-138, 1994

5. W.B. Kleijn, J. Haagen " A speech coder based on decomposition of characteristics waveforms" IEEE Proc ICASSP95, pp. 508-511, 1995,.

6. X.Q. Sun, F. Plante, B.M.G. Cheetham, W.T.K. Wong, "Phase modelling of speech excitation for low bit-rate sinusoidal transform coding" IEEE ICASSP'97, Vol.3 pp.1691-1694, Munich, April 1997.

7. S.D. Watson, B.M.G. Cheetham, W.T.K. Wong, P.A. Barrett, A.V. Lewis, "A voice activity detector for the ITU-T 8kbits/s speech coding standard G.729", Eurospeech'97, pp.1571-1574, Rhodes, September 1997.

8. E. Zwicker, "Subdivision of the audible frequency range into critical bands", J. Acoust. Soc. Am., Vol.33, pp.244-255, 1961

9. D. Marston "Gender adapted speech coding", IEEE ICASSP'98, Vol.1 pp.357-360, Seattle, May 1998.