Phonetically Switched Tree coding of speech with a G.727 Code Generator

Pravin Ramadas and Jerry D. Gibson
Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA
Email: pravin_ramadas@umail.ucsb.edu, gibson@ece.ucsb.edu

***Abstract* - A simple phonetic classification method based on frame error energy level in combination with the AMR-VAD1 Voice Activity Detection algorithm is used to classify input speech into five different modes: Voiced, Onset, Unvoiced, Hangover and Silence. Each mode is coded at a suitable bit-rate using a Tree Coder with perceptual error weighting criteria and a G.727 Code Generator. Apart from efficient coding, the Tree coder smoothes transitions between different coding modes. At an average bit-rate just greater than 16 kbps the phonetically switched Tree coder produces speech quality equivalent to the G.727 ADPCM coder at 32 kbps for nearly a 50% bit-rate saving. Apart from bit-rate savings; bit-rate scalability, moderate delay, and good tandeming performance are also achieved.**

I. INTRODUCTION

G.727 is an ITU-T standard embedded Adaptive Differential Pulse Code Modulation (ADPCM) speech coder, standardized for digital telephony applications. The G.727 coder has an embedded quantization structure that offers a bit-rate scalability option [1].

G.727 is a waveform coder with desirable properties for speech coding, such as high quality, low delay, low complexity and good tandeming performance, but these properties come at the cost of high bit-rate. In this paper we propose a speech coding method which classifies speech into different phonetic modes and codes each mode suitably using a Tree coder with G.727 code Generator to achieve the desired properties of a G.727 ADPCM coder at a lower bit-rate.

Phonetic classification of speech has been used to achieve low bit-rate coding of speech in [2-4]. Phonetic classification identifies the phonetic modes in speech in order to code each mode with just a sufficient number of bits to reduce the overall bit-rate. Our previous work, on the phonetically switched ADPCM speech coder in [2] codes each phonetic mode in speech differently at an appropriate bit-rate with G.726 ADPCM coder to achieve speech quality comparable to G.726 ADPCM at 24 kbps but at an average bit-rate less than 16 kbps. The coder had a delay of 40 ms due to the chosen phonetic classification procedure.

This research has been supported by NSF Grant Nos. CCF-0728646 and CCF-0917230.

In the codec proposed here, we use a modified phonetic classification procedure based on frame error energy to reduce the delay and complexity. Different modes in speech are coded at an appropriate bit-rate using a tree coder with perceptual error weighting criteria. The tree coder not only effectively codes speech samples but also helps in smoothing transitions between different modes of speech by the virtue of its look-ahead. A comfort noise generation procedure is used to improve the perceptual output quality during silence. These improvements result in speech quality equivalent to 32 kbps G.727 at an average bit-rate of about 16 kbps for 55% Voice Activity sequences.

The paper is organized as follows. Section II describes the phonetic classification procedure. Section III describes the mode based tree coder components, and Section IV gives an overview of the comfort noise generation method. Section V presents the speech quality and bit-rate improvements achieved by this coder in comparison with a Phonetically Switched (PS)-ADPCM coder and a standard G.727 coder at 32 kbps. Section VI discusses the tandem performance of the new PS-Tree coder.

II. PHONETIC CLASSIFICATION PROCEDURE

The phonetic classification method for the codec described in this paper uses AMR-VAD1 Voice Activity Detection in combination with a frame error energy based phonetic classification method to classify speech into five different modes; Voiced, Onsets, UnVoiced, Hangover and Silence.

Speech samples are grouped into frames of 90 samples each and coded with G.727 ADPCM at 16 kbps. The error energy is computed from the difference between the original and reconstructed speech for the frame. The frame error energy is compared to the threshold obtained using a weighted combination of the error energies of the previous speech frames and average Voiced and Unvoiced error energies, to make a Voiced/Unvoiced decision. The AMR-VAD1 algorithm is used for Voice/Silence

classification and the phonetic classification method further classifies the Voice frames into Voiced and Unvoiced modes. Note that there is mis-alignment between AMR-VAD1, which makes the VAD decision on 160 speech samples, and the phonetic classification method, which makes a mode decision on 90 sample frames. In such cases, if the previous VAD decision is Voice, then the current frame is also considered Voice and fed into the phonetic classification method for further mode classification. If the previous VAD decision is silence, the current frame energy is compared with the frame energy of the previous silence frames for Voice/Silence decision. The complete phonetic classification procedure explained above is shown in Figure 1.

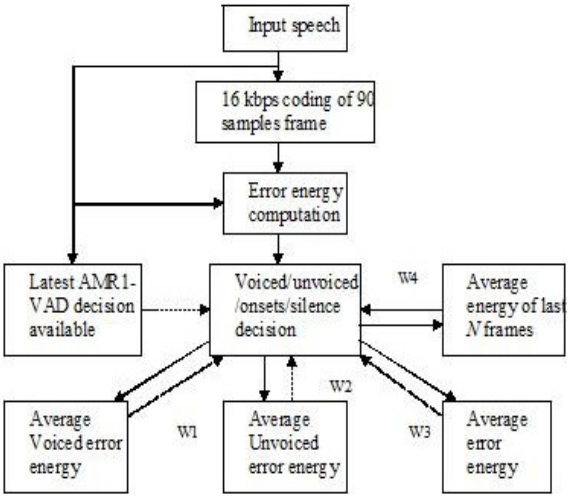

Figure 1. Phonetic Classification procedure for silence, Voiced, Unvoiced mode classification. w's are the weights given to each parameter to form the decision threshold.

This phonetic classification method classifies higher reconstruction error regions as Voiced, which are coded at high bit-rate and lower reconstruction error regions as Unvoiced, which are coded at low bit-rate. This method ensures that different regions in speech are coded at sufficient bit-rate and the over all reconstruction error is reduced. Onset mode frame is the first Voiced frame that follows Unvoiced or Silence frame. First five Silence frames are classified as Hangover mode in order to provide smooth transition from Voice into Silence. This phonetic classification method reduces computational complexity and encoder delay to 12.375 ms. Figure 2 shows the phonetic classification result for a sample speech sequence.

Header bits are attached before each frame to identify the mode type at the decoder. In our implementation, Onsets and Voiced frames are coded at 32 kbps and Unvoiced and Hangover frames are coded at 16 kbps and the Silence frames are coded at 1.5 kbps with both Silence Descriptor frames (SID) and Silence no update frames (SNU) as explained in Section IV. A 2-bit frame header is used to identify these frames.

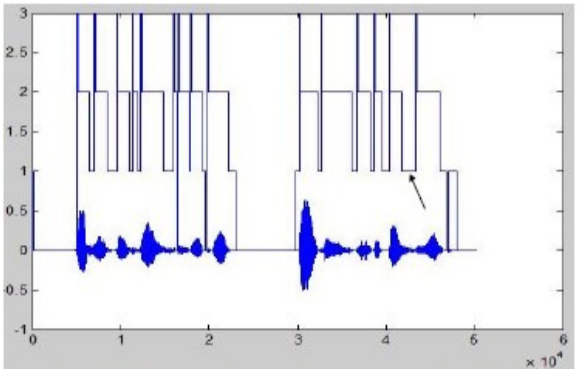

Figure 2. Phonetic Classification result of the sequence 'Acid burns holes in old cloth'. Segments enclosed with values 3, 2, 1, 0 are onsets, voiced, unvoiced and silence respectively. Region marked by the arrow is slight mis-classification of Voiced mode as Unvoiced but output quality is not affected since this region is sufficiently coded.

III. MODE BASED G.727 TREE CODING

Tree coding is a multi-path search procedure to encode each speech sample based on best long term fit to the input waveform. Tree coder encodes each input sample at time instant *k,* using only the data at times $j \leq k$. Tree coders improve on this approach by delaying the encoding decision by a few more samples, say *L*, such that the input samples at time instants $j \leq k + L$ are used to encode the sample at time instant *k*. By this delayed decision, the tree coder searches the most likely paths among the 2^L possible paths to find the best fit for the current sample [5,6]. The fit and the consequent path selection are based on a suitable error measure.

The design of a tree coder consists of selecting a code generator, a tree search algorithm, a distortion measure and a path map symbol release rule as shown in Figure 3. The tree search, in combination with the code generator and appropriate distortion calculation method, chooses the best candidate path to encode the current input sample s(k). The Symbol release rule decides on the symbol(s) to encode in order to reconstruct the sample at the decoder.

The part of the G.727 ADPCM encoder which emulates the decoder is the code generator. It generates candidate outputs *s'(k)* for the given path map. The *M-L* Tree search algorithm is used to reduce the computational complexity by limiting the multi-path search to M most likely paths rather than all 2^L possible paths. *M=10* and *L=10* are used in this encoder implementation. Perceptual weighted error is used as the distortion measure.

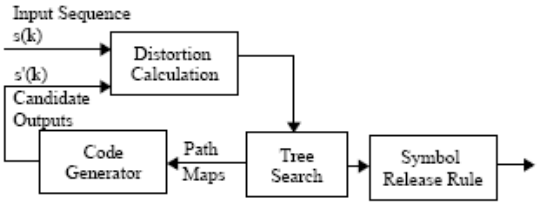

Figure 3. Block diagram of Tree coder

This criterion helps in choosing the path where the noise is masked by the speech spectrum. The perceptually weighted distortion is obtained by filtering the reconstruction errors along the depth-L path through the Perceptual error weighting filter as given [6]:

$$W(z) = \frac{1 - \sum_{i=1}^{N} a_i z^{-i}}{1 - \sum_{i=1}^{N} \mu^i a_i z^{-i}}$$

where the value of μ is 0.86. a_i 's are the short term predictor coefficients calculated from the current speech frame. The value of N is 5. The single symbol release rule is used as path map symbol release rule.

The phonetically-switched tree coder with the G.727 code generator (PS-Tree coder) is shown in Figure 4. Voiced and Onset frames are coded at 32 kbps while Unvoiced and hangover frames are coded at 16 kbps, with this PS-Tree coder maintaining a single set of state parameters across modes. Since the tree coder looks ahead into the future L samples to code the current sample, it helps to smooth transitions between different modes. Silence frames are encoded using Comfort Noise parameters, which are explained in detail in Section IV. In the bit packing step, the frames are packed with appropriate header bits for the decoder to identify the mode information.

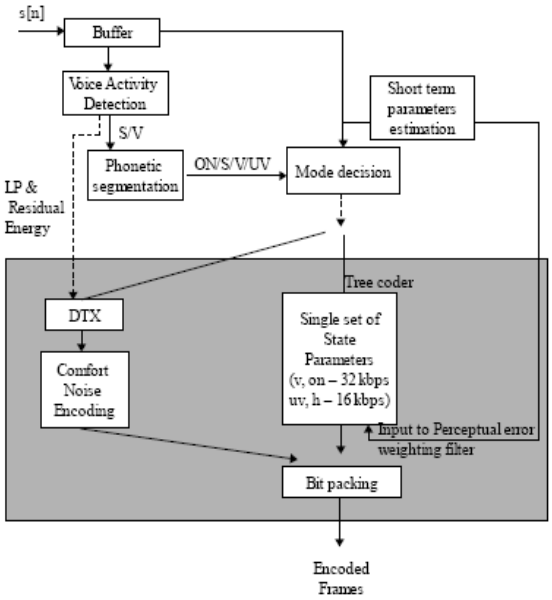

Figure 4. Phonetically-Switched Tree coder with G.727 code generator

The Phonetically-Switched G.727 decoder is shown in figure 5. Based on the mode information decoded from the header bits of the frame, G.727 ADPCM decoder operates at appropriate bit-rate to decode speech. Silence frames are reconstructed using Comfort Noise Generation procedure.

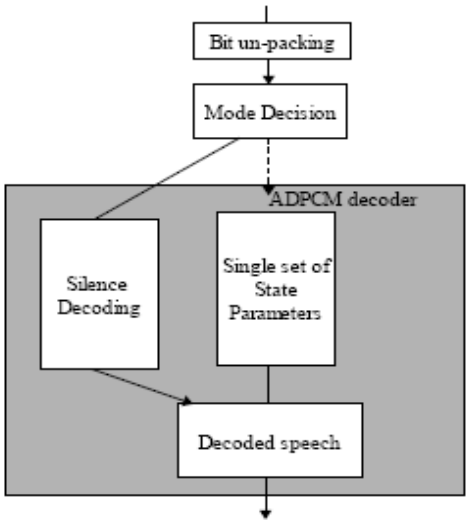

Figure 5. Phonetically-Switched G.727 decoder

IV. COMFORT NOISE GENERATION

The comfort Noise Generation procedure improves the perceptual quality during silence frames. The proposed CNG procedure is based on G.711 Appendix II [7] and G.729B CNG [8]. A Discontinuous transmission (DTX) scheme similar to G.729B evaluates the background and whenever there is significant change in spectral and energy content, Silence information is transmitted using the Silence Descriptor (SID) frame. A minimum spacing of three frames is imposed between consecutive SID frames to save bandwidth during non-stationary noise. When the DTX does not indicate a significant change, a Silence No Update (SNU) frame is transmitted instead of SID frame. During SNU frame the previously received SID information is used.

The autocorrelation values of the previous six frames are averaged, and LP coefficients are calculated. These are converted into PARCORs rather than LSFs for computational simplicity [7], if any of the PARCOR exceeds absolute value of 1.0 then the LP filter is unstable and previous SID frame information is used. Each LP coefficient is represented by a 6 bit value. The residual energy is scalar Quantized using a 5-bit nonuniform quantizer [8]. The resulting bit-rate for encoding silence is about 1.2 kbps. At the decoder, LP coefficients are obtained from the PARCORs. A white Gaussian

excitation is stored in the decoder which is scaled according to the residual energy and filtered through the LP synthesis filter to generate comfort noise. The LP synthesis filter coefficients are interpolated with previous LP synthesis filter coefficients to minimize the effect of spectral distortion due to quantization of PARCORs.

V. RESULTS

In this section, the Phonetically-Switched Tree coder with the G.727 code generator is compared with G.727 ADPCM at 32 kbps using PESQ-MOS values. The sentences used in the experiments are:

1. "Acid burns holes in old cloth. Fairy tales are fun to read"
2. "Oak is strong and also gives shade"
3. "A lathe is a big tool"
4. "Wipe the grease off your dirty face"

The first two sequences are male and the second two are female. The sequences used in this experiment have about 55% Voice Activity and are clean without any background noise in Table 1 and with background noise in Table 2.

Table 1 compares the PESQ-MOS values of PS-Tree coder with G.727 at 32 kbps for clean speech sequences. The PS-Tree coder produces speech quality comparable to 32 kbps G.727 coder but at an average bit-rate of about 16 kbps. Clearly there is almost 50% savings in bit-rate by using the PS-Tree coder at a moderate encoder delay of 12.375 ms.

TABLE 1
Comparison of PESQ-MOS values of G.727 at 32 kbps with PS-Tree coder for clean speech sequences.

Clean Sequence	G.727 at 32 kbps - PESQ	PS-Tree coder (Av. Bit-rate: 16 kbps) - PESQ
Fairytales	3.933	3.923
Oak	3.879	4.096
Lathe	3.917	3.887
Wipeface	3.940	3.906

TABLE 2
Comparison of PESQ-MOS values of G.727 at 32 kbps with PS-Tree coder for noisy speech sequences.

Noisy Sequence	G.727 at 32 kbps - PESQ	PS-Tree coder (Av. Bit-rate: greater than 16 kbps) - PESQ
Fairytales	3.225	3.219
Oak	3.746	3.856
Lathe	3.501	3.473
Wipeface	3.323	3.332

Table 2 compares the PESQ-MOS values of PS-Tree coder with G.727 at 32 kbps for noisy speech. Again, the G.727 PS-Tree Coder performs similar to G.727 at 32 kbps.

VI. TANDEM PERFORMANCE

As a result of the heterogeneous networking environment with each network likely to use different speech codecs, it is important to make sure the end-to-end speech quality is not affected significantly due to the asynchronous tandem operation of different speech codecs. The degradation is particularly due to transcoding at network interfaces and distortion accumulation due to repeated coding [9]. To ensure that the improved PS-Tree coder maintains acceptable tandeming with some commonly used narrow-band speech codecs such as AMR-NB (at 12.2 kbps) and G.729 (at 8 kbps), tandem experiments are performed with those codecs and the results are compared with the tandem performance of G.727 at 32 kbps. All inputs are clean speech.

Table 3 represents the PESQ-MOS results of the tandem performance of G.727 with AMR-NB and G.729. Table 4 represents the PESQ-MOS results of the tandem performance of PS-Tree coder with AMR-NB and G.729. The first row of the tables show the order of tandeming. The coder mentioned first is used in the first stage and the coder following it is used in the second stage.

TABLE 3
Tandeming performance of G.727 at 32 kbps, measured in PESQ-MOS. First Row shows the order of tandeming. X-Y, coder X is used in the first stage and Y in the second stage

Sequence	**G727-AMR**	**G727-G729**	**G727-G727**	**AMR-G727**	**G729-G727**
Fairy	4.103	3.792	3.726	3.782	3.642
oak	3.812	3.581	3.885	3.785	3.599
lathe	3.976	3.456	3.789	3.766	3.538
wipe	4.052	3.809	3.768	3.886	3.720

TABLE 4
Tandeming performance of PS-Tree coder (PS), measure in PESQ-MOS. First Row shows the order of tandeming. X-Y, coder X is used in the first stage and Y in the second stage

Sequence	**PS-AMR**	**PS-G729**	**PS-PS**	**AMR-PS**	**G729-PS**
Fairy	3.981	3.755	3.932	3.673	3.576
oak	4.056	3.781	4.096	3.798	3.601
lathe	3.754	3.564	3.887	3.786	3.498
wipe	3.935	3.713	3.800	3.843	3.650

From Tables 3 and 4, it can be seen that the self-tandeming performance of the PS-Tree coder is better than G.727. This is because noise is introduced in the silence part of the original clean sequence in the G.727 coder while silence is preserved by the PS-Tree coder using Silence encoding. G.727 is a waveform following coder and is known to have good tandeming performance. By comparing Tables 3 and 4, we see that the PS-Tree coder tandeming performance is very close to the performance of G.727 at 32 kbps and results in good speech quality. Hence the PS-Tree coder has a good tandeming performance.

VII. CONCLUSIONS

The proposed Phonetically-Switched Tree coder with the G.727 Code Generator achieves speech quality equivalent to 32 kbps G.727 ADPCM at an average bit-rate of about 16 kbps for conversational speech. The coder uses a simple phonetic classification method that reduces computational complexity and encoder delay to moderate 12.375 ms. This fully-backward adaptive speech coder has the option of bit-rate scalability and also tandems well with other popular narrowband speech coders.

VIII. REFERENCES

[1] ITU-T Recommendation G.727, "5-, 4-, 3- and 2-bits sample embedded Adaptive Differential Pulse Code Modulation (ADPCM)", 1990

[2] P. Ramadas and J. D. Gibson, "A phonetically switched ADPCM speech coder", in *Proceedings*, 42nd Asilomar Conference on Signals, Systems and Computers , Oct. 26 - 29, 2008

[3] S. Wang and A. Gersho, "Phonetically-based vector excitation coding of speech at 3.6 kbit/s," in *Proceedings*, IEEE ICASSP, May 1989.

[4] S. Wang and A. Gersho, "Improved Phonetically-Segmented Vector Excitation Coding at 3.4 Kb/s," in *Proceedings*, IEEE ICASSP, San Francisco, pp. 349-352, March 1992.

[5] N. S. Jayant and S. A. Christensen, "Tree-Encoding of speech using the (M, L)- Algorithm and Adaptive Quantization", *IEEE Trans. on Communications*, Vol. COM-26, NO. 9, September 1978.

[6] A.C. Goris and J. D. Gibson, "Incremental Tree coding of Speech," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 511-516, July 1981.

[7] ITU-T G.711 Appendix II: A Comfort noise payload definition for ITU-T G.711 use in packet-based multimedia communication systems

[8] ITU-T Recommendation G.729, Coding of speech at 8 kbits/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)

[9] J. D. Gibson and B. Wei, "Tandem Voice Communications: Digital Cellular, VoIP, and Voice over Wi-Fi", IEEE Communications Society, Globecom 2004.

[10] H. C. Woo and J. D. Gibson, "Low-Delay Tree Coding of Speech at 8 Kbps," *IEEE Trans. Speech and Audio Proc.*, pp. 361-370, July 1994.