A Phonetically Switched ADPCM speech coder

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

Abstract

Voice activity detection and phonetically-based segmentation are used to classify input speech into four modes: onset, silence, unvoiced and voiced. Each phonetic segment is coded at a suitable bitrate depending on the mode type, using a G.726 ADPCM encoder and preserving distinct encoder state information for each mode. The proposed speech coder achieves PESQ-MOS equivalent to G.726 ADPCM at 24 kbps but at an average rate less than 16 kbps while encoding a typical telephone conversation. A moderate 40 ms encoder delay is incurred .

I. Introduction

G.726 is an ITU-T standard that uses Adaptive Differential Pulse Code Modulation (ADPCM) waveform encoding for digital telephony at bit rates between 16 and 40 kbps.

This paper presents a phonetically switched G.726 ADPCM encoder that can give high quality at a lower bitrate. Based on phonetic information, the speech segments are classified into onset, silence, unvoiced and voiced modes. Each mode type is treated distinctly and separately encoded at a suitable bitrate.

We use Voice Activity Detection (VAD) to remove silence prior to coding. Unvoiced speech is coded with just two bits per sample (at 16 kbps), while voiced speech and onsets are encoded at four bits per sample (at 32 kbps).

The paper is organized as follows. Section II explains the method used for phonetic segmentation of speech. Section III describes the mode-based encoder based on G.726. Section IV presents the spectrogram comparison of input and output waveforms. PESQ-MOS values comparing with 24 kbps G.726 output are also presented.

This research has been supported by the California Micro Program, Applied Signal Technology, Cisco, Dolby Labs, Inc., Qualcomm Inc, and Sony-Ericsson, and by NSF Grant Nos. CCF-0429884, CNS-0435527, and CCF-0728646.

II. Phonetic segmentation for mode classification

Phonetic information has been used very effectively for speech coding at low bitrates [1,2]. We draw on that work and other related work on Voice Activity detection and segmentation [3-5].

Silence is a significant portion of a telephone conversation. Therefore, it is effective to detect silence and remove it prior to coding. A robust Voice activity detection (VAD) algorithm helps to detect silence/background from speech even under noisy conditions. Reference [4] compares the performance of recent ITU-T and ETSI VAD algorithms based on objective, psychoacoustic and subjective parameters. Among all those VAD algorithms, AMR1 VAD is more robust and performs well under noisy conditions[4]. Hence AMR1 VAD has been chosen for silence/background detection. The algorithm computes SNR in nine bands and the decision is based on a comparison between the SNRs and a threshold, which is different in each band[5].

AMR1 VAD operates on speech frame of length 160 samples. After the frame has been detected as either silence/background or voice activity region, a phonetically-based segmentation algorithm[3] is used to further classify the voice activity region into voiced and unvoiced. The algorithm uses a weighted linear combination of seven feature values extracted from the speech segment and the resulting value is thresholded to obtain a one bit parameter for voiced/unvoiced decision. The set of features are: zero-crossing rate, low-band speech energy, first reflection coefficient, preemphasized energy ratio, second reflection coefficient, forward pitch prediction gain and backward pitch prediction gain.

The phonetically-based segmentation algorithm takes 90 sample (11.25 ms) speech frames. Hence the AMR1-VAD decision is segregated into 90 sample speech segments from 160 samples in order to feed the segmentation algorithm. If there is a conflict in segregation between voice and silence, a voice decision is taken. The segmentation algorithm then classifies the 90 sample speech segments into

voiced and unvoiced modes. Fig.1 shows an example of the speech segmentation procedure.

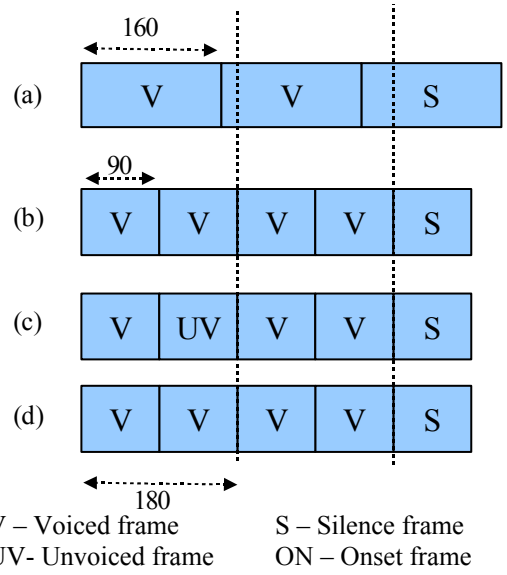

Fig 1. (a) AMR1-VAD decision on 160 samples (b) AMR1-VAD decision is segregated into 90 sample speech segments to feed the phonetic-segmentation algorithm – notice that in the fourth frame, conflict between voiced and silence decision is resolved by taking a voiced decision (c) voiced/unvoiced decision taken in voice activity region by phonetic segmentation algorithm (d) Final mode decision after 3-pt median smoothing and marking onsets. Assume the previous speech segment as voiced.

After the phonetic segmentation algorithm classifies the segments as voiced or unvoiced. The voiced segments that immediately follow unvoiced segments are classified as onsets. After that, a 3-point median smoothing is applied to remove any spurious mode decision. The smoother requires previous, current, and the next segment in order to make a smoothing decision. Hence a 180 sample buffer for current and next speech segments is needed for the mode decision. This implies that two AMR1 VAD decisions need to be buffered. This introduces a delay of 40 ms in the encoder.

III. G.726 coding based on mode decision

After the mode decision is made as described in the previous section, voiced/unvoiced/onset speech segments are encoded separately. The encoder block diagram is shown in Fig 3. Speech samples in each segment are encoded using the appropriate mode based encoder depending on the mode decision. By this scheme each mode type uses different ADPCM parameters. The ADPCM state parameters that are preserved uniquely for each mode are: step size, quantizer scale factor adaption, adaptation speed control, adaptive predictor and tone transition detector. Since G.726 is backward adaptive, this information does not have to be transmitted to the decoder.

The unvoiced mode is encoded at 16 kbps. Voiced and onsets are encoded at 32 kbps. The encoded bits are packed with a start bit for each frame (90 samples encoded in each frame). This start bit before each frame helps the decoder to identify the mode type of the frame and hence decode it appropriately. An example of generating frames with start bit is shown in Fig. 2. If the start bit is '0' the mode of the current frame is the same as the previous frame. If the start bit is '1' then a mode change is indicated and a complimentary bit is used to identify the mode type of the current frame. Silence frames are not encoded. The decoder inserts zeros on silence mode information.

Previous Frame → / Complimentary bit	**ON**	**S**	**UV**	**V**
0	*V*	*UV*	*ON*	*S*
1	*V*	*ON*	*S*	*UV*

Table 1. Complimentary bit for each mode type when the start bit is '1' indicating a mode change

The decoder block diagram is shown in Fig 4. At the decoder, the appropriate mode based decoder is chosen based on the start bits present before each encoded speech frame. If the start bit indicates a silence frame then the decoder just inserts zeros for 90 samples. Due to mode switching between speech segments there are noisy samples introduced in the transition boundaries at the encoder.

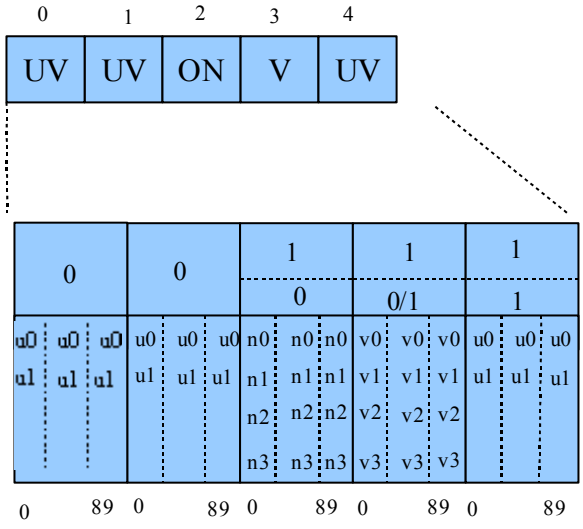

Fig 2. A sequence of five frames generated at the encoder. Each frame contains 90 samples. Unvoiced frames contain two bits per sample. Voiced and onset frames have four bits per sample.

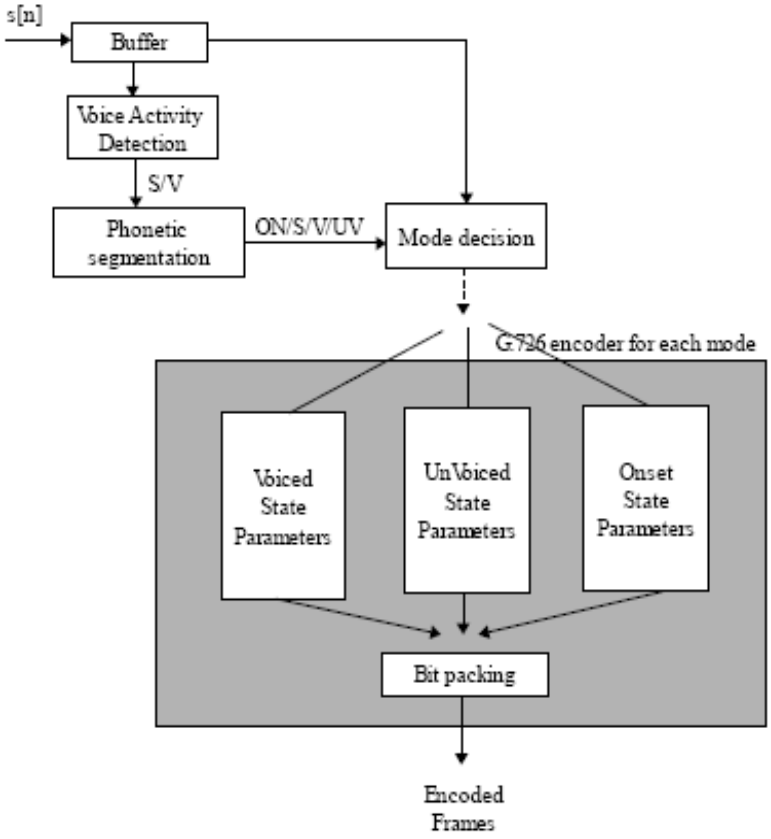

Fig.3 G.726 ADPCM encoder based on phonetically-based mode classification

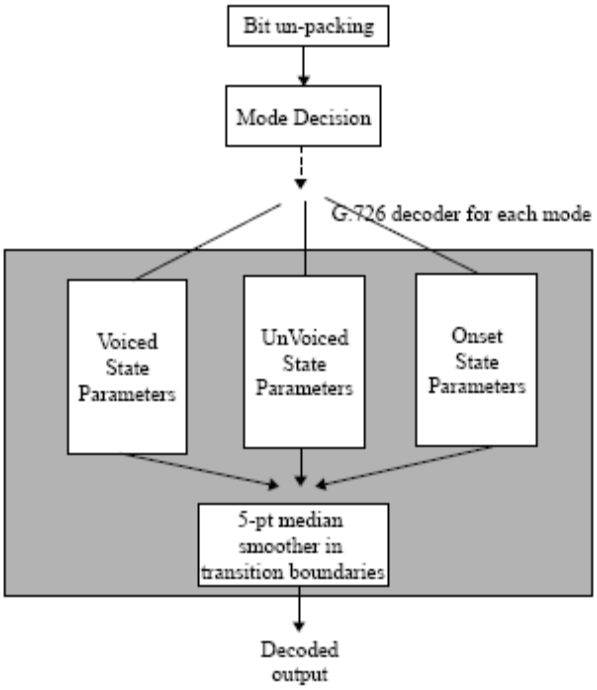

Fig.4 G.726 ADPCM decoder based on mode informaiton

These noisy samples are removed by applying a 5-point median smoothing in the transition boundaries. Improvement in speech quality due to median smoothing at the transition boundaries can be clearly seen from the spectrogram shown on the last page of this paper.

IV. Results

Mode decisions obtained on a sample sentence of speech using method described in Section II is shown in Fig. 5.

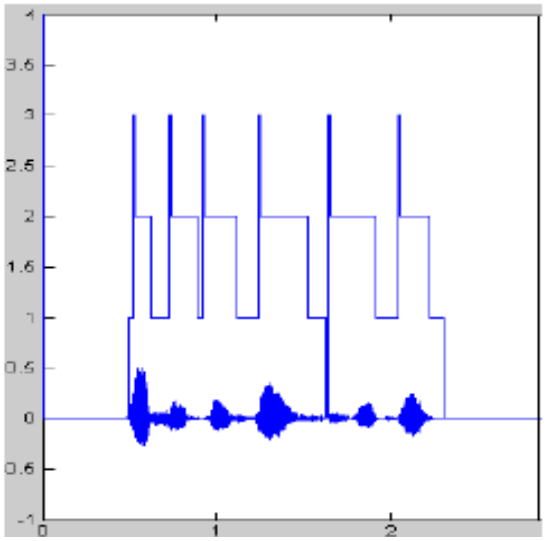

Fig.5 "Acid burns holes in old cloth." Segments enclosed with value 3 are onsets, 2 are voiced, 1 are unvoiced, 0 are silence.

Spectrograms comparing the input speech sample, decoded speech sample without 5-point median filtering and decoded speech sample with 5-point median filtering are shown on the last page of this paper.

In a typical telephone conversation, silence is present at nearly 50% of the time. The remaining time is occupied by voiced, unvoiced and onset segments. Since unvoiced is encoded at 16 kbps, and voiced and onsets are encoded at 32 kbps, the resulting average bit rate is less than 16 kbps. PESQ-MOS values in Table 2 show that the quality of the G.726 with mode classification based encoding is similar to G.726 at 24 kbps. This shows the significant gain in bitrate achieved using mode classification based encoding while maintaining high quality.

List of sentences used in Table 2.
1. "Acid burns holes in old cloth"
2. "Oak is strong and also gives shade"
3. "A speedy man can beat this track mark"
4. "Wipe the grease off your dirty face"

Sentences	G.726 ADPCM at 24 kbps	G.726 ADPCM with Mode based encoding
1	3.449	3.483
2	3.702	3.837
3	3.502	3.392
4	3.533	3.568

Table 2. PESQ-MOS values of G.726 24 kbps Vs G.726 with mode classification based encoding

V. Conclusion

A novel speech waveform encoding method based on phonetically segmenting the speech waveform into onset, silence, unvoiced and voiced and encoding each mode separately at a suitable bitrate has been presented. While encoding a typical telephone conversation, silence occupies nearly 50% of the samples. The remaining 50% of the samples are onsets, voiced or unvoiced. Zeros are inserted in the place of silence, the unvoiced mode is encoded at 16 kbps, and the voiced and onset modes are encoded at 32 kbps. This results in an average bitrate less than 16 kbps and quality comparable to 24 kbps G.726. Quality can be further improved by inserting comfort noise instead of zeros during silence.

VII. References

[1] Shihua Wang and A. Gersho, "Improved Phonetically- Segmented Vector Excitation Coding at 3.4 Kb/s," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, San Francisco, vol. 1, pp. 349-352, March 1992.

[2] Wang, S. and Gersho, A., "Phonetically-based vector excitation coding of speech at 3.6 kbit/s," *Proc. IEEE Intern. Conference on Acoustics, Speech, and Signal Processing,* Glasgow, May 1989.

[3] J.P. Campbell, Jr. and T.E. Tremain, "Voiced/Unvoiced Classification of Speech with Applications to the US. Government LPC-IOE Algorithm," *Proceedings* of *IEEE International Conference on Acoustics, Speech. and Signal Processing.* vol. 1, pp. 437-476, Tokyo, April 1986.

[4] Beritelli.F; Casale.S; Ruggeri.G; Serrano.S, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors", Signal Processing Letters, IEEE ,Vol. 9 , Issue 3 , March 2002, pp.85 - 88

[5] 3GPP TS 26.094 v7.0.0 (2007-06). Technical report on Mandatory speech codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Voice Activity Detector (VAD)

[6] G.726 technical specification: ITU-T Software Tool Library 2005 User's manual.

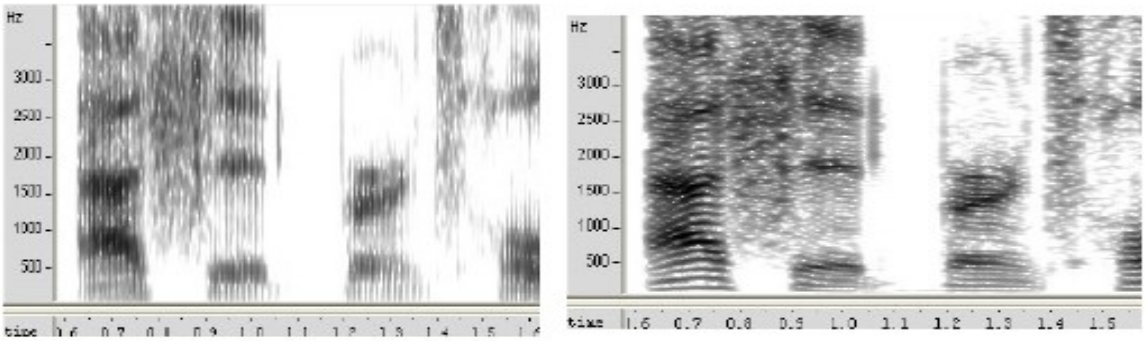

Fig.6 Wideband and Narrowband spectrograms of Original sequence "Acid burns"

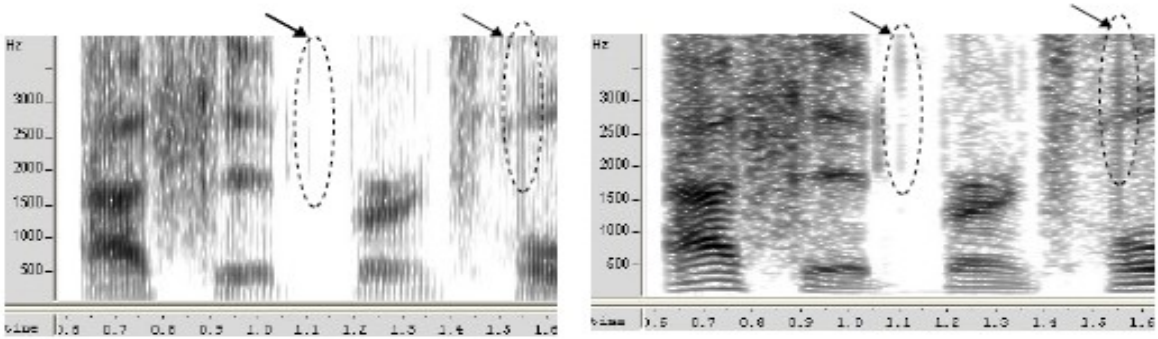

Fig.7 Wideband and Narrowband spectrograms of G.726 ADPCM mode classification based generated output without 5-pt median smoothing in the mode transition boundaries. Pointed and encircled regions indicate noise in the transition boundaries.

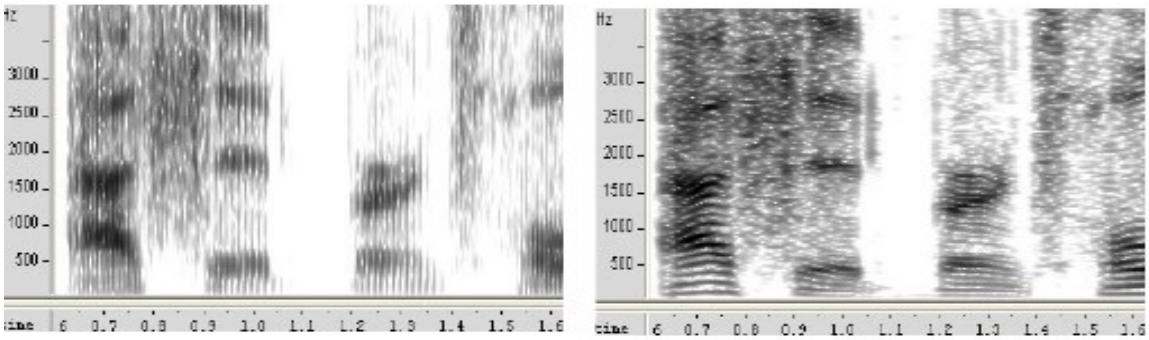

Fig.8 Wideband and Narrowband spectrograms of G.726 ADPCM mode classification based generated output with 5-pt median smoothing in the mode transition boundaries. Noise indicated in the Pointed and encircled regions in Fig. 7 have been removed.