# IMPLEMENTATION OF AN ENHANCED FIXED POINT VARIABLE BIT-RATE MELP VOCODER ON TMS320C549

*A.E. Ertan, E.B. Aksu, H.G. İlk, H. Karcı, Ö. Karpat, T. Kolçak, L. Şendur, M. Demirekler, A.E. Çetin*

TUBITAK-BILTEN / Speech Processing Laboratory
Middle East Technical University
Electrical and Electronics Engineering Department, D Block, 06531
Ankara - TURKEY
speech@bilten.metu.edu.tr

**ABSTRACT**

In this paper, a fixed point Variable Bit-Rate (VBR) Mixed Excitation Linear Predictive Coding ($MELP^{TM}$) vocoder is presented. The VBR-MELP vocoder is also implemented on a TMS320C54x and it achieves virtually indistinguishable federal standard MELP quality at bit-rates between 1.0 to 1.6 kb/s. The backbone of VBR-MELP vocoder is similar to that of federal standard MELP. It utilizes a novel sub-band based voice activity detector in the back-end of encoder to discriminate background noise from speech activity. Since proposed detector uses only parameters extracted in the encoder, its computational complexity is very low.

## 1. INTRODUCTION

Designing high quality vocoders at low bit-rates ($< 4.0$ kb/s) has been the focal point of considerable research activity in the last decade [1, 2]. The quest for those efficient speech coding algorithms lead to new types of "low bit-rate coders" which are important for enhancing secure communications in government, military and civil applications. The applications of these speech codecs are numerous.

Code Excited Linear Predictive (CELP) coders can produce high quality speech at bit-rates above 4.0 kb/s. However a rapid detoriation in the decoded speech quality is observed when bit-rate is decreased below 4.0 kb/s. The main reason for this detoriation is the small number of bits which are not sufficient for a waveform matching operation in the time domain. Recently, several approaches have been attemped for encoding speech at lower bit-rates. These coders can be divided into three major groups:

1. **Prototype Interpolation Coders**: Prototype Waveform Interpolation (PWI) [3] and its variants such as Characteristic Waveform (CW) representation and Time-Frequency Interpolation (TFI) [4].
2. **Harmonic Coders**: Sinusoidal Transform Coders (STC) [5] and Multi-band Excitation Vocoders (MBE) [6].
3. **LPC Based Vocoders**: Mixed Excitation Linear Prediction (MELP) coder [7].

U.S. Deparment of Defense, Digital Voice Processing Consortium (DoD-DDVPC) selected a MELP based vocoder [7] as the recommended new federal standard in 1996. This new standard provides equal or improved performance over 4.8 kb/s FS1016 CELP coder at only 2.4 kb/s. In fact an increasing demand for digital speech coding applications made further development of MELP and other low bit rate speech coding systems inevitable.

Our research group designed and developed a VBR-MELP vocoder on a fixed point TMS320C54x DSP processor. Real-time implementation of fixed point VBR-MELP vocoder requires artihmetic operations, like calculation of square root, sine, cosine and logarithm, to be done using look-up tables and proper interpolation techniques. In some cases, 32 bit arithmetic, instead of 16, is required when precision is an important design factor. Since VBR-MELP vocoder demands huge processing power due to its complex nature, the 100 MIPS version of TMS320C549 is selected for the DSP processor. The first version of the VBR-MELP vocoder is presented in [8]. This paper however descibes a more complex VAD which shows better noise immunity.

## 2. VARIABLE BIT-RATE MELP VOCODER

In low bit-rate speech coders, the main objective is to synthesize perceptually high quality speech with a minimum set of parameters, which can then be efficiently coded at bit-rates below 4.0 kb/s. MELP algorithm's synthesizer's speech quality is exceptionally good at 2.4 kb/s. However, original MELP coder operates at 54 bits/22.5 ms frames at all times regardless of the information content of the speech signal. If different parts of speech can be represented with different number of bits required for the sufficent parameter set, the average bit-rate can be reduced down to 1.2 kb/s without any compromise in the decoded output speech quality.

Fixed-rate 2.4 kb/s MELP vocoder separates bit-stream format into two sections whose bit allocation tables can be found in [7]:

1. Mixed or pulse excited parts for voiced sections.
2. Noise excited parts for unvoiced sections.

Synthesis of unvoiced sections do not require *Fourier magnitudes*, *pitch period*, *bandpass voicing decisions* and *aperiodic flag*. Hence, these bits are used for error protection (13 bits/frame) and transmission of sequence type (7 bits/frame in the place of *pitch period*). Since a header is always transmitted in a variable bit-rate vocoder, the requirement for transmission of pitch period information is eliminated for unvoiced frames. Furthermore, error protection is unnecessary in variable bit-rate systems, so there is no need to allocate bits for this purpose. In addition, gain is generally stable in unvoiced sections, therefore transmission of the

This work was supported by ASELSAN Inc.

first gain parameter is also redundant. As a result, elimination of transmission of these parameters decreases the required bits from 54 bits/frame to 30 bits/frame for unvoiced frames.

In addition to the bit-rate reduction in unvoiced sections, further bit-rate reduction can be achieved by efficient coding of silence and background noise sections of the input sequence, which covers nearly 40 percent of a typical conversation. For this purpose, the new VBR-MELP coder utilizes a voice activity detector to detect silent regions and makes efficient coding to reduce the average bit-rate.

In this novel variable rate MELP coder, a two-bit header is used to classify frame type: Voiced, unvoiced and silence/noise. For the silence/noise frames, the parameters of the first frame of the silence regions are transmitted as if it is an unvoiced frame and these parameters are repeatedly used until a header showing different frame type other than silence/noise is encountered. Table 1 shows final bit allocation.

Table 1: Bit allocation table for variable bit-rate MELP vocoder.

| Parameters | Voiced | Unvoiced | Noise |
|---|---|---|---|
| Header | 2 | 2 | 2 |
| LSFs | 25 | 25 | – |
| Fourier Magnitudes | 8 | – | – |
| Gain (2 per frame) | 8 | 5 | – |
| Pitch | 7 | – | – |
| Bandpass Voicing | 4 | – | – |
| Aperiodic Flag | 1 | – | – |
| Total Bits / Frame | 55 | 32 | 2 |

The proposed VAD is robust in noisy environments, such as vehicle noise [9]. The details of the VAD algorithm are given in the next section.

## 3. VOICE ACTIVITY DETECTOR

Voice activity detector designed in this work is a combination of two detectors: First one utilizes a finite state machine for the detection of silence parts [10]. This detector uses energy of the frame, ratio of energies in consecutive frames and zero-crossing number as the feature set. Due to the selection of the parameter set, this detector does not have any noise robustness. The second VAD is the one used in Pan-European Digital Cellular Mobile Telephone Service (GSM-VAD) [11]. It utilizes two seperate detectors, one for making voice activity decision by comparing some parameters with adaptive thresholds and one for adapting these thresholds.

MELP algorithm extracts following parameters within the encoder.

- Decomposition of signal into 5 subbands.
- Pitch period.
- Bandpass voicing strenghts.
- LSFs.
- Two gain calculation for the first and the second half of the frame.

In order to decrease computational complexity, these parameters are used directly within our VAD.

The new proposed VAD uses the distance measure (1), $D_k$, based on the logarithm of the ratio of differences between signal and noise energies in the subbands to the variance of noise in the corresponding bands, which is also used in [12].

$$D_k = 10 \cdot \log \left[ \frac{1}{L} \sum_{l=1}^{L} \frac{(E_l^k - \mu_l)^2}{\sigma_l^2} \right] \qquad (1)$$

$E_l^k$ is the energy parameter of the $k^{th}$ frame for the $l^{th}$ subband over a time window and computed by (2). $\mu_l$ and $\sigma_l$ are the estimated mean and variance of the background noise in the $l^{th}$ subband, respectively.

$$E_l^k = \frac{1}{N_l} \sum_{n=1}^{N_l} (s_l(n))^2 \qquad l = 1, 2, \cdots, L \qquad (2)$$

$N_l$ is the number of samples in the $l^{th}$ subband.

In our system, there is no decimation in subband decomposition. $N_l$ is equal to 90 samples corresponding to half of an 22.5 ms MELP frame. Furthemore, since signal is decomposed into 5 subbands, $L$ is selected to be 5.

The original method is reported to be successful in end-point detection in noisy environments. However, end-point detection system described in [12] assumes that the initial few frames are always noise and therefore the noise variances in the subbands are extracted from these frames. Unfortunately, in a typical telephone conversation, the first few frames may contain speech information. Therefore, it is impossible to make this kind of assumption in our system. To overcome this problem, the required variances of the noise in the subbands are extracted by a similiar method described in GSM-VAD.

The block diagram of our VAD is illustrated in Figure 1. The system has two main parts labeled as '*VAD1*' and '*VAD2*':

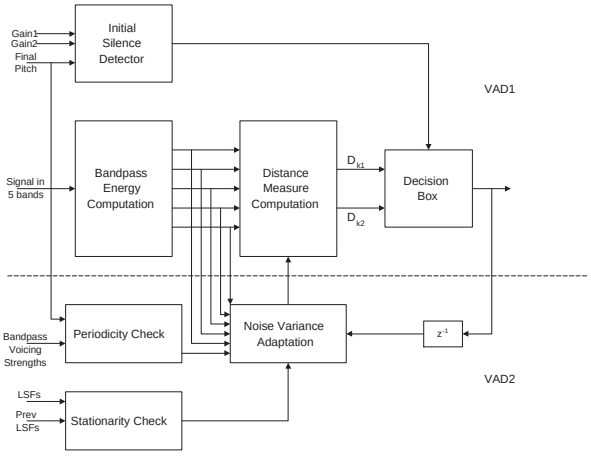


Figure 1: Voice Activity Detector for VBR-MELP Vocoder.

'*VAD1*' is used to detect the presence of the speech signal. First, energy in 5 subbands are extracted for the first and the second half of the frame and $D_k$ is computed twice for the two sub-frames to obtain $D_{k_1}$ and $D_{k_2}$. Besides, an initial silence detector is used to detect inaudible signal by comparing the gain values with an experimentally derived threshold. Furthermore, this detector also provides echo-suppression in some degree when no background noise is present: Voiced sections and information tones always

have high energies. If a strong periodic structure is detected with a gain smaller than a second threshold, that frame is assumed to belong to a part of an echo signal. The output of this initial silence detector and computed values of $D_k$ are used by the decision system to make final decision about the voicing state. The decision box contains a finite state machine,which consists of four different states about silence detection:

1. Silence (SI): These regions contain only background noise.
2. Primary Detection of Signal (PD): These regions are primary detection for signal which may contain information. Note that if the system stays in this stage for longer than a pre-determined duration, 'Speech Enable' state is activated.
3. Speech Enable (SE): These regions contain speech signal.
4. Hang-over Period (HO): Silence is detected in these regions, however, to eliminate misclassification of the weak fricatives and the final nasals at the end of the speech, a hang-over period is inserted in the system.

State transitions are performed according to the values of the two coefficients:

- $PDF_k$ : Primary detection of speech for the $k^{th}$ frame. Its value is set to 0, if silence is declared in initial silence detector or $D_{k_n}$ is below $D_{t1}$. Otherwise, it is set to 1.
- $SDF_k$ : Definite presence of speech for the $k^{th}$ frame. Its value is set to 1, if $D_{k_n}$ exceeds $D_{t2}$. Otherwise, it is set to 0.

Values of $D_{t1}$ and $D_{t2}$ are obtained experimentally. These values are selected such that clipping of speech and misclassification of silence frames are minimized. The system is found to be optimum, when $D_{t1}$ and $D_{t2}$ are set to 5 and 10, respectively.

The state transition diagram is given in Figure 2. Transitions from *PD* to *SE* and *HO* to *SI* requires some past information, i.e. memory. To go from *PD* to *SE*, $PDF$ must be set to 1 for 20 half-frames, that corresponds to a wait state of 225 ms. Transfer from *HO* to *SI* in general requires 45 ms in our system. This relatively short hang-over period is due to the robustness of our system to the background noise. Furthermore, if the *talk-spurt* duration is shorter than 56.25 ms, this period is also reduced to 22.5 ms. In simulation studies, it is observed that only *stop-gaps* are missed with this approach.

States are updated twice in a 22.5 ms frame, one for the value of $D_{k_1}$ and one for the value of $D_{k_2}$.

'*VAD2*' is used in parallel with '*VAD1*' to update the variances of noise in 5 subbands. Note that since speech signal is first high-pass filtered with cut-off frequency of 60 Hz [7], $\mu_l$'s for all bands are zero. The adaptation is performed by '*Noise Variance Adaptation*' block, which takes the required parameters from '*Stationarity Check*' block, '*Periodicity Check*' block and the decision of '*VAD1*' of the previous frame. Periodicity control is performed by comparing first bandpass voicing strength with an expermentally derived threshold. If this threshold is exceeded and one or more bandpass voicing strength other than first one are equal to 1, frame is concluded to be periodic. Non-stationarity detection is performed by a novel algorithm based on comparison of the peaks of the spectrum estimated from LSFs and the difference between the consecutive LSFs of current and previous frames. This new method can obtain location of the peaks within 25 Hz error range with 95 percent accuracy and provides reliable detection of spectrum changes. The details of the algorithm can be found in [9]. Note that proposed non-stationarity detector has one frame delay.

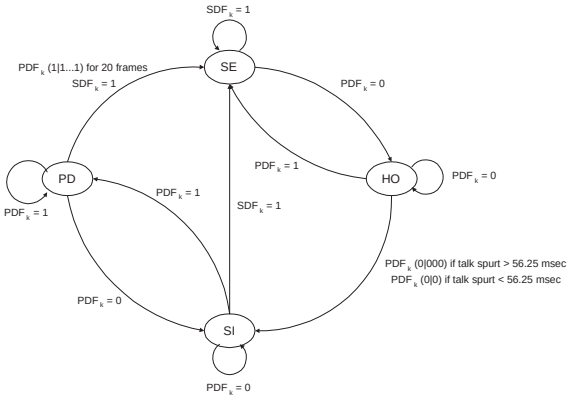


Figure 2: State transition diagram of the decision box. SI stands for silence state. PD stands for primary detection state. SE stands for speech detected frames. HO stands for hangover state.

In order to update noise variance in subbands, following conditions must be met:

1. Signal must be stationary for a period of time longer than $S_x \cdot 22.5$ ms. $S_x$ is the number of frames for the system to wait before adaptation takes place in which signal is stationary.
2. Signal must not be periodic. Since information tones has long duration, they may be classified as long stationary regions. These regions must not be included in noise adaptation.
3. Decision state of final decision box in '*VAD1*' must be same for previous two frames.

If these conditions are met, variances of noise for 5 subbands are calculated and then continuously averaged in every frame using (3) until at least one of the three conditions given above is violated.

$$\sigma_l^2 = \frac{\sigma_l^{2'} \cdot N_a + E_l^k}{N_a + 1} \tag{3}$$

$N_a$ are the number of stationary frames after the first frame of adaptation and $\sigma_l^{2'}$ is the variance of noise in the previous frame.

Value of $S_x$ can be varied from 6 to 20. It is observed that 8 is a reasonable value, since practically none of the unvoiced phonemes last longer than 180 ms.

'*VAD2*' also has one frame delay. Frame length of this block is 180 sample as in MELP vocoder [7].

## 4. VBR-MELP IMPLEMENTATION

The whole MELP coder was implemented in both fixed point and floating point on PC with *C* programming languauge and performance tests of both implementations are conducted on this platform. In the initial fixed point implementation, the quality of the decoded speech was not satisfactory, even unacceptable for some speakers. It was found out that the main problem was the precision, occurred for some specific type of input sequences. To overcome this problem, LPC filter's coeffcients calculation, corresponding LSFs calculations and interpolation functions are performed in 32 bit arithmetic instead of 16 bit, in order to obtain outputs as close as

possible to the floating point implementation. After achieving desired quality, the whole vocoder is written in TMS320C54x assembler, since it is experimented that the only C compiler available for this processor is not capable of generating highly optimized codes required for our purposes. After this step, by proper adjustments in the bitstream generation and utilization of the voice activity detector, a variable bit-rate version of this vocoder is implemented in the same way.

## 5. SIMULATION RESULTS

Tests are conducted only to obtain the performance of the voice activity detector in car environment for various SNR levels. In these tests, SNR values are calculated from the portions which includes speech signal only. Percentage of clipped regions and misclassified noise regions are tabulated in Table 2. Test sequence is obtained from a 50 sec telephone conversation[1]. 57 percent of the conversation consists of only background noise. A Volvo340 car noise[2], driven on a rainy asphalt road is added to the clean speech in various levels to obtain desired SNR levels.

Table 2: Performance of proposed VAD in various SNR levels. $P_{cl}$ stands for the percentage of clipped regions with respect to the overall speech sections. $P_{ms}$ stands for the percentage of the missed regions with respect to the background noise sections.

| SNR level(dB) | $P_{cl}$ | $P_{ms}$ | Avg. bit-rate |
|---|---|---|---|
| $\infty$ | 3.84 | 3.76 | 977.83 bps |
| 30 | 2.29 | 3.03 | 940.05 bps |
| 20 | 2.35 | 13.23 | 1076.00 bps |
| 15 | 2.41 | 20.36 | 1088.50 bps |
| 10 | 1.83 | 10.81 | 1012.00 bps |
| 5 | 9.92 | 19.09 | 998.00 bps |

From these experiments, it is observed that our VAD works satisfactorily when SNR of the sequence is higher than 10 dB. The clipped regions in these levels are occurred due to a long laugh, in which detector assumes these regions as noise and equate noise variances to the energy of the speech in that region. Therefore, some utterances are missed due to this wrong adaptation. However, with the beginning of background noise, system adapts thresholds again and recovers itself. The misclassified silence regions are mostly due to the hang-over period and little energy variations in background noise. Finally, it can be seen that nearly in all cases, average bit-rate is around 1000 bps which makes further $1:2.4$ compression over fixed-rate MELP vocoder without a considerable loss of quality.

## 6. CONCLUSION

In this paper, we have presented the implementation of a new variable rate fixed point MELP vocoder. This vocoder achieves significant bit-rate reduction, from 2.4 kb/s to an average of 1.2 kb/s, over federal standard MELP vocoder with similar output quality. This vocoder manages to obtain these lower rates by assigning small number of bits for unvoiced and background noise detected frames. Furthermore, a novel voice activity detector is presented, which utilizes a distance measure based on sub-band energies of input signal and energy estimation of noise. This VAD is observed to provide reliable discrimination of background noise and speech portions even in low SNR values. Finally, it can be stated that since only three frame types are present in our system, a fourth one can also be defined to encode different kind of speech signal, like onsets in the beginning of the utterance.

[1] This conversation is taken from 'Switchboard Corpus - Recorded Telephone Conversations.' database collected by Texas Instruments.

[2] Institute for Perception, TNO, The Netherlands.

## 7. REFERENCES


[1] A. Gersho. Advances in speech coding and audio compression. *Proceedings of IEEE, vol. 82*, 1994.

[2] A.S. Spanias. Speech coding: A tutorail review. *Proceedings of IEEE, vol. 82*, 1994.

[3] W.B. Kleijn. Continuous representations in linear predictive coding. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 201–204, 1991.

[4] Y. Shoham. High-quality speech coding at 2.4 to 4.0 kbps based on time-frequency interpolation. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages II167–II170, 1993.

[5] R.J. McAulay and T.F. Quatieri. *Advances in Speech Processing*, chapter Low Rate Speech Coding Based on Sinusoidal Model. Marcel Dekker, 1991.

[6] D.W. Griffin and J.S. Lim. Multiband excitation vocoder. *IEEE Trans. Acoust., Speech, and Signal Process., vol. 36*, pages 1223–1235, 1988.

[7] L.M. Supplee, R.P. Cohn, J.S. Collura, and A.V. McCree. Melp: The new federal standard at 2400 bps. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1591–1594, 1997.

[8] E.B. Aksu, A.E. Ertan, H.G. İlk, H. Karcı, Ö. Karpat, T. Kolçak, L. Şendur, M. Demirekler and A.E. Çetin. Implementation of a variable bit-rate melp vocoder on TMS320C548. *2nd European DSP Education Conference*, 1998. To be published.

[9] A. Erdem Ertan. Spectrum non-stationarity detection algorithm based on line spectrum frequencies and related applications. Master's thesis, Bilkent University, 1998.

[10] Y. Yatsuzuka. Highly sensitive speech detector and high-speed voiceband data discriminator in DSI-ADPCM systems. *IEEE Trans. on Communications, vol. 30*, pages 739–750, 1982.

[11] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd. The voice activity detector for the pan-european digital cellular mobile telephone service. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 369–372, 1989.

[12] Engin Erzin. *New Methods for Robust Speech Recognition*. PhD thesis, Bilkent University, 1995.