Performance Comparison of Linear Prediction based Vocoders in Linux Platform

Lani Rachel Mathew*, Ancy S. Anselam† and Sakuntala S. Pillai‡
Department of Electronics and Communication Engineering
Mar Baselios College of Engineering and Technology, Nalanchira
Thiruvananthapuram 695 015, Kerala, India
*M.Tech Student, lanirachel@gmail.com
†Assistant Professor, ancy_anselam@yahoo.co.in
‡Dean (R & D), sakuntala.pillai@gmail.com

***Abstract*—Linear predictive coders form an important class of speech coders. This paper describes the software level implementation of linear prediction based vocoders, viz. Code Excited Linear Prediction (CELP), Low-Delay CELP (LD-CELP) and Mixed Excitation Linear Prediction (MELP) at bit rates of 4.8 kb/s, 16 kb/s and 2.4 kb/s respectively. The C programs of the vocoders have been compiled and executed in Linux platform. Subjective testing with the help of Mean Opinion Score test has been performed. Waveform analysis has been done using Praat and Adobe Audition software. The results show that MELP and CELP produce comparable quality while the quality of LD-CELP coder is much higher, at the expense of higher bit rate.**

Keywords - Vocoder, linear prediction, code excited, low delay, mixed excitation, CELP, LD-CELP, MELP, Praat, Linux

I. Introduction

Speech coding is the encoding of speech signals to enable transmission at bit rates lower than that of the original digitized speech. The human auditory system can capture only certain aspects of a speech signal. Thus, perceptually relevant information of a speech signal can be extracted to produce an equivalent-sounding wave at a much lower bandwidth.

Linear prediction is a widely used compression technique in which past samples of a signal are stored and used to predict the next sample [1]. In the basic linear predictive coder prototype, prediction is done over a time interval of one pitch period using adaptive linear delay and gain factors. This basic prototype produces intelligible but artificial-sounding speech output, and various techniques have been researched to improve the perceptual quality.

Variants to the linear prediction coders are Code Excited Linear Prediction (CELP) and Low-Delay CELP (LD-CELP) which use forward and backward linear prediction respectively, along with the *codebooks*, i.e. lookup tables with codevectors corresponding to speech residual signals [2], [3]. In Mixed Excitation Linear Prediction (MELP) [4], an additional classification of speech is introduced - the jittery voiced speech. Mixed excitation, i.e the mixing of periodic and noise excitation, is another distinguishing feature of the MELP model.

This paper aims at comparing the three types of linear prediction based vocoders in terms of their bit rate and perceptual quality. In comparing vocoders, subjective testing of the voice quality is a major step in the evaluation of a vocoder. A vocoder will finally be accepted in the market only if humans are satisfied with the voice quality. Keeping this fact in consideration, subjective evaluation using the Mean Opinion Score (MOS) test has been conducted.

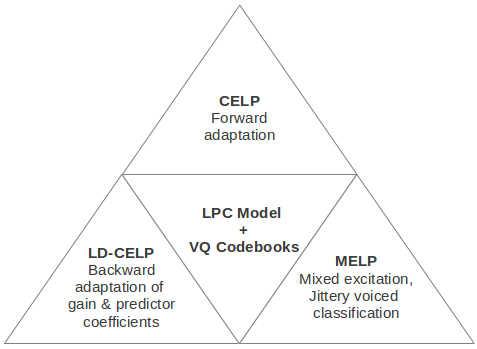


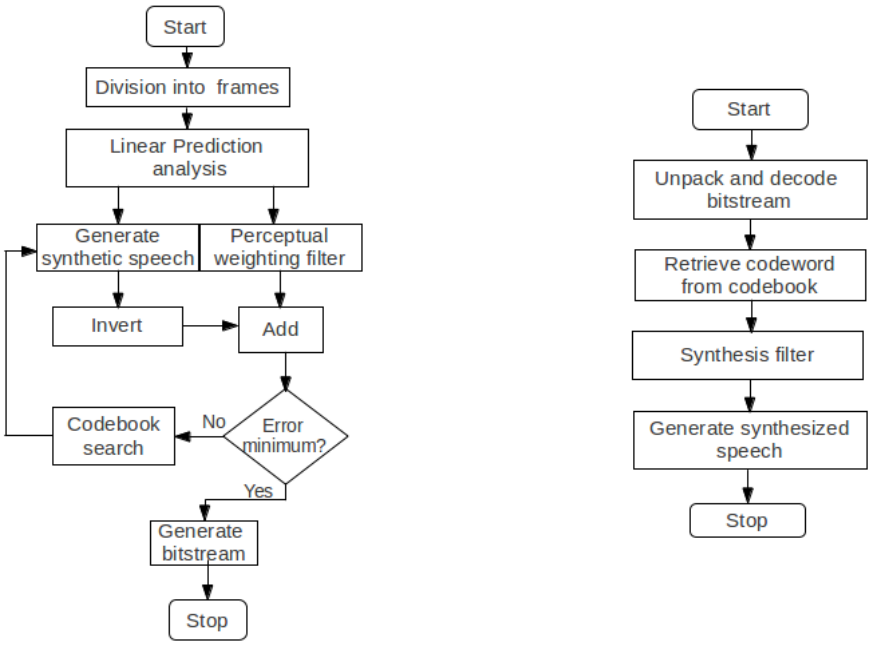
Fig. 1. LPC model as the core of CELP, LD-CELP and MELP algorithms

The paper is organized as follows: Section II gives a brief overview of the vocoder algorithms. The bit allocation and bit rate calculations are also described. Section III describes the method adopted in implementing and testing the vocoder. The results obtained and their implications are discussed in Section IV, followed by concluding remarks.

II. Vocoder Algorithms

The Linear Prediction model forms the core of the CELP, LD-CELP and MELP algorithms, as depicted in Fig. 1. In the LPC model, speech signals are classified into voiced and unvoiced signals. Voiced signals are generated when the vocal cords vibrate and are represented in the LPC source-filter model as periodic excitation. Unvoiced signals are generated by turbulence of air in the vocal tract and do not involve the vocal cords. These signals are usually represented as white Gaussian noise.

The LPC coder consists of a linear predictor having adaptive delay and gain factors [1]. Since there are sounds in speech that are produced by a combination of voiced and unvoiced signals, it has been observed that important perceptual speech



(a) Encoder (b) Decoder

Fig. 2. Generalized Design Flow for Linear Prediction based Vocoder

information gets missed out from the predictor output. Also a slight error in the predictor coefficients will lead to more speech information being missed out. This unaccounted residual output of the LPC coder contains important data on how the sound signal is perceived by the human auditory system.

A. Code Excited Linear Prediction Algorithm

It was proposed in [2] that the prediction residual signal could be used to enhance the perceptual quality of the coder output. A codebook containing a list of codevectors is searched to obtain a closest match to the residual signal. The index of the codevector is selected such that minimization of the perceptually weighted error metric is obtained. The codec earned its name due to the use of codebooks to obtain *codes* for modeling the speech signal. The same codebook is available at both the transmitter and the receiver. At the receiver, the index is used to obtain the codevector and use it in the synthesis filters.

CELP uses the Analysis-by-Synthesis (AbS) method in which the transmitter *analyzes* the signal to produce linear prediction coefficients, and then uses these coefficients to *synthesize* the speech signal within the transmitter itself. An error signal is generated and codevectors are selected from the codebook in order to minimize the perceptually weighted mean square error.

In the CELP Transmitter, the transmitter first splits the input speech into frames of around 30 ms. Short-term linear prediction is performed, i.e. formants (peaks of the spectral envelope) are estimated for each input frame. The transfer function of the perceptual weighting filter of CELP is given by the following equation.

$$W(z) = \frac{1 - Q(z)}{1 - Q(\frac{z}{\gamma})} \tag{1}$$

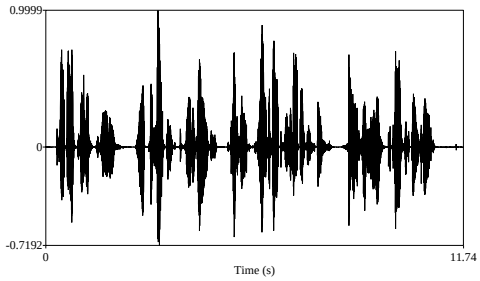
where

$$Q(z) = \sum_{i=1}^{M} q_i z^{-i} \tag{2}$$

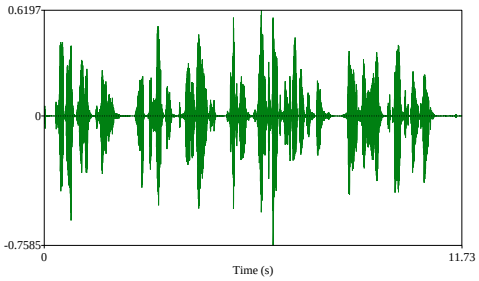
$$Q(\frac{z}{\gamma}) = \sum_{i=1}^{M} \gamma^i q_i z^{-i}, 0 < \gamma < 1 \tag{3}$$

where M is the LPC predictor order and q_i's are the quantized LPC coefficients. After finding the short term LPC coefficients, each frame is split into four subframes, i.e. 7.5 ms each, which are given as input to the long-term prediction filter. The pitch and the intensity of the speech signal are estimated. An optional post filtering stage may be added after decoding to enhance the quality of the output signal. The CELP bit allocation [7] is shown below:

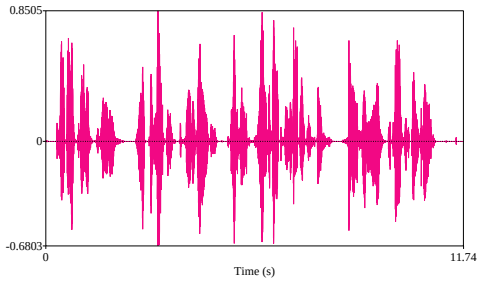
Parameter	No./frame	Total bits/frame
Linear prediction coefficients	10	34
Pitch period	4	28
Adaptive codebook gain	4	20
Stochastic codebook index	4	36
Stochastic codebook gain	4	20
Synchronization	1	1
Error correction	4	4
Future expansion	1	1
Total bits/30 ms frame		144



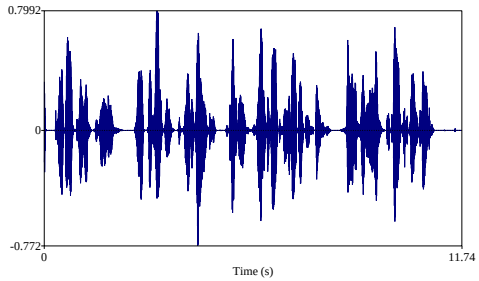
(a) Original Sound File



(b) CELP synthesized output



(c) LD-CELP synthesized output



(d) MELP synthesized output

Fig. 3. Time domain representation

The 30 ms frame of CELP corresponds to 240 samples for a sampling rate of 8000 Hz. Thus the bit rate of CELP is 144/30ms = 4.8 kb/s.

B. Low Delay Code Excited Linear Prediction Algorithm

The CELP and LD-CELP algorithms differ only in the type of adaptation - forward and backward respectively - in which linear prediction is performed. Low delay is achieved by the use of a backward-adaptive predictor and short excitation vectors (5 samples each) [3]. Only the index of the excitation codebook is transmitted - all other parameters are updated by backward adaptation of previously quantized speech. LD-CELP uses a modified system function for the weighting filter as given below.

$$W(z) = \frac{1 - Q(\frac{z}{\gamma_1})}{1 - Q(\frac{z}{\gamma_2})}, 0 < \gamma_2 < \gamma_1 \leq 1 \qquad (4)$$

where the parameters γ_1 and γ_2 are tuned to optimize the quality of the coded speech and Q(z) is given by the expression given below.

The LD-CELP bit allocation is shown below [8]:

Parameter	No. of Bits
Excitation Index	7
Excitation Gain	2
Sign codebook gain	1
Total bits/2.5 ms vector	10
Total bits/10 ms frame	40

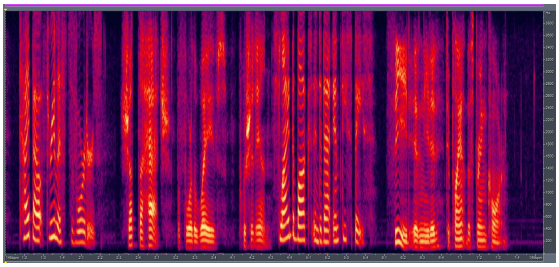
5 samples, i.e. a vector corresponds to 0.625 ms for a sampling rate of 8000 Hz. Thus the bit rate of LD-CELP is 10/0.625ms = 16 kb/s.

C. Mixed Excitation Linear Prediction Algorithm

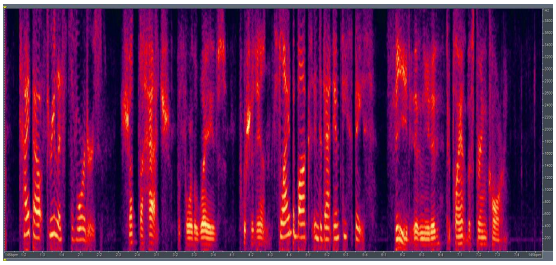
In the MELP coder, there are three classifications for the speech signal - voiced, unvoiced and jittery voiced. The third classification is done when voicing transitions occur, i.e. when aperiodic but not completely random excitations occur. Another feature is that the shape of the excitation pulse is also extracted from the input signal. Pulse shaping filters and noise shaping filters are used to filter the pulse train and white noise excitations. 'Mixed excitation' refers to the total excitation signal which is the sum of the filtered output periodic and noise excitations. The MELP bit allocation is described in the table below [5]:

Parameter	Voiced	Unvoiced
LSF parameters	25	25
Fourier magnitudes	8	-
Gain (2 per frame)	8	8
Pitch, overall voicing	7	7
Bandpass voicing	4	-
Aperiodic flag	1	-
Error protection	-	13
Sync bit	1	1
Total bits/22.5 ms frame	54	54

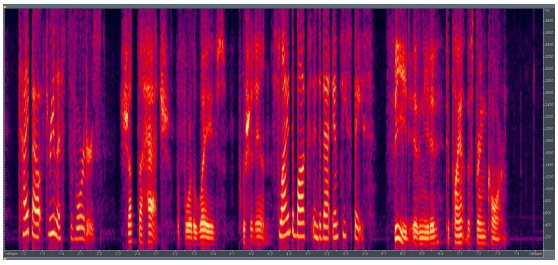
The 22.5 ms frame of MELP corresponds to 180 samples for a sampling rate of 8000 Hz. Thus the bit rate of MELP is 54/22.5ms = 2.4 kb/s.



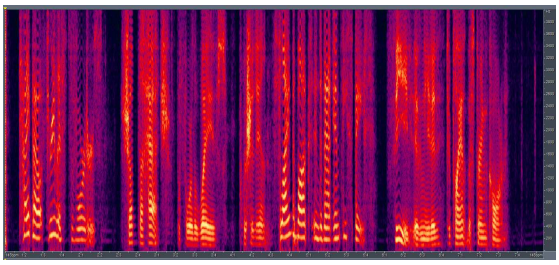
(a) Original Spectrogram



(b) CELP Spectrogram



(c) LD-CELP Spectrogram



(d) MELP Spectrogram

Fig. 4. Spectrograms of speech signals

III. Method

The C programs of the vocoders were compiled with GCC(GNU Compiler Collection) and built using the GNU Make utility in a Linux platform. For waveform conversions, Sound eXchange (SoX) software was used. SoX is an open-source tool for speech file manipulations. The analysis and synthesis commands used in each vocoder are enlisted below:

1) CELP commands
   a) *Analysis*: ./celp -i inputfile.wav -o outputfile
   b) *Synthesis*: ./celp -c outputfile.chan -o outputsynth
   c) *Copy spd (speech data) file to raw file*: cp outputsynth.spd outputsynth.raw
   d) *Convert to wav file*: sox -r 8000 -b 16 -c 1 -e signed-integer outputsynth.raw outputsynth.wav
   e) *Playing the file*: padsp play outputsynth.wav
2) LD-CELP commands
   a) *Analysis*: ./ccelp inputfile.wav encoderout.out
   b) *Synthesis*: ./dcelp encoderout.out outputsynth.raw
3) MELP commands
   a) *Analysis*: ./melp -a -i inputfile.wav -o encoderout.out
   b) *Synthesis*: ./melp -s -i encoderout.out -o outputsynth.raw

In LD-CELP and MELP, conversion of headerless raw format to wav file format and playing of the output file are performed in the same manner as that of CELP.

A. Waveform analysis

Waveform analysis was performed using Praat, a tool used for phonetic analysis of speech. A standard speech sample *source.wav* was used for waveform analysis. Time domain representation of the speech files as well as pitch and intensity waveforms were plotted using the Praat Objects and Picture windows. Spectrogram analysis was done with the help of Adobe Audition software.

B. Subjective Testing: Mean Opinion Score

Evaluation of the perceptual quality was done using the Mean Opinion Score (MOS) test. Due to time constraints, informal testing of the codecs was conducted with 10 evaluators. Speech samples recorded in the English language were given as input to the vocoders. Three samples were recorded in a quiet environment, while two speech samples were recorded with loud background music. Logitech h110 stereo headsets were used for voice recording and playback. The evaluators were given an initial training on the MOS test, and their scores were recorded. MOS scores and their interpretations are given below [8].

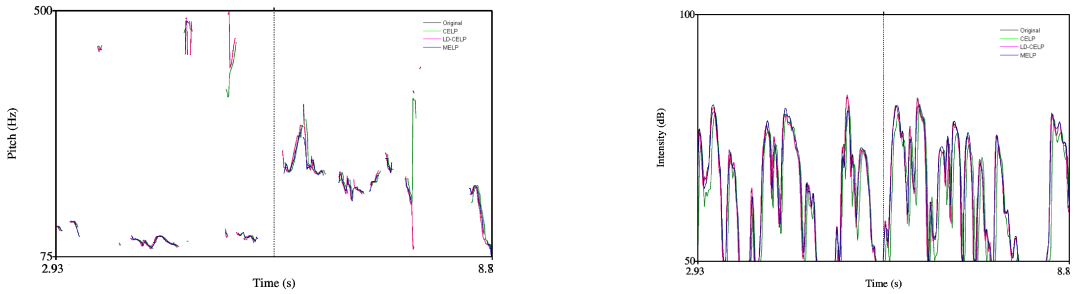
MOS	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

A MOS score of 4 or 5 indicates toll quality speech while a score of 1 or 2 indicates synthetic speech.

IV. Results and Discussion

A. Waveform analysis

The results of waveform analysis using Praat software are shown in Figs. 3 to 5. Fig. 3 shows the time domain representation of the original and synthesized speech files of CELP, LD-CELP and MELP coders. Fig. 4 depicts the spectrograms of the original and synthesized speech waveforms. Fig. 5 shows the pitch and intensity contours.



(a) Comparison of pitch contours

(b) Comparison of intensity contours

Fig. 5. Pitch and intensity waveforms

TABLE I
INPUT SPEECH FILE DETAILS

Filename	File Size (kB)	Duration (s)
male_eng.wav	170	10.65
female_eng.wav	176	10.98
male_fem_conversation.wav	319	19.91
male_noisy_eng.wav	447	27.93
female_noisy_eng.wav	856	53.49

TABLE II
MOS SCORE FOR VOCODERS

Filename	CELP	LD-CELP	MELP
male_eng.wav	3.10	4.06	2.82
female_eng.wav	3.24	4.02	2.76
male_fem_conversation.wav	3.10	3.76	3.12
male_noisy_eng.wav	3.00	4.58	3.24
female_noisy_eng.wav	2.52	4.26	1.72
Average MOS	**2.992**	**4.136**	**2.732**

From the time domain waveforms, it can be concluded that the overall shape of the original wave has been preserved. However peaks have been clipped at certain portions which result in decrease in clarity of the speech output. The spectrograms show the frequency content of the speech waveforms as a function of time. The pitch and intensity graphs also show slight variations in the output of the vocoders. Reliable estimates of the perceptual quality can be made only by conducting subjective tests using human listeners.

B. Subjective Testing: Mean Opinion Score

All recorded input speech files were sampled at a rate of 8 kHz. The details of speech input files are shown in Table I. The average input speech file size was 393.6kB and average duration 24.59s. The MOS scores corresponding to each input speech file and the average MOS score obtained for each vocoder are shown in Table II.

The MOS scores in Table II show that LD-CELP has the highest perceptual quality (toll quality) among the three vocoders. The perceptual quality of CELP and MELP vocoders is rated less, with CELP scoring slightly higher than MELP.

V. CONCLUSION

The results of the comparison indicate that a choice can be made only based on the application of the vocoder. In applications where the focus is on low delay and high perceptual quality, as in two-way communication systems, the LD-CELP algorithm at 16 kb/s is the ideal candidate. In areas where low bit rate is essential, MELP is the best candidate because it can work at bit rates as low as 2.4 kb/s and gives intelligible output. When both low bit rate and good quality are required, the CELP coder at 4.8 kb/s seems to be the most suitable coder. In this study, the number of evaluators for the MOS test was limited to 10 due to time constraints. For more accurate results, the number of evaluators needs to be increased.

ACKNOWLEDGMENT

The authors would like to thank the students and faculty of the department for providing speech samples and for their participation in the Mean Opinion Score testing. Special thanks goes to Karthika Balan for her valuable help in the MOS testing process.

REFERENCES

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, 1975.

[2] M. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at at very low bit rates," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 10, pp. 937–940, 1985.

[3] Juin-Hwey Chen, R. V. Cox, Y. C. Lin, N. Jayant and M. J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE Journal on Selected Areas in Communications*, vol. 10, issue 5, pp. 830–849, 1992.

[4] Alan McCree, Kwan Truong, E. B. George, T. P. Barnwell and V. Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new U.S. Federal Standard," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 200–203, 1996.

[5] Lynn M. Supplee, Ronald P. Cohn, John S. Collura and Alan V. McCree,"MELP: the new Federal Standard at 2400 bps," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1591–1594, 1997.

[6] L. Mathew, A. Anselam and S. S. Pillai, "Analysis of LD-CELP coder output with Sound eXchange and Praat software," unpublished.

[7] Wai C. Chu, "Speech Coding Algorithms: Foundation and Evolution of Standardized Coders," Wiley Interscience, 2003.

[8] Olivier Hersent, Jean-Pierre Petit and David Gurle, "Beyond VoIP Protocols: Understanding Voice Technology and Networking Techniques for IP Telephony," John Wiley & Sons, 2005.