



## Low Bit Rate Speech Coding

Jaspreet Singh<sup>1</sup>, Mayank Kumar<sup>2</sup>

<sup>1</sup>*Asst. Prof.ECE, RIMT Bareilly,*

<sup>2</sup>*Asst. Prof.ECE, RIMT Bareilly*

### ABSTRACT

*Despite enormous advances in digital communication, the voice is still the primary tool with which people exchange ideas. However, uncompressed digital speech tends to require prohibitively high data rates (upward of 64kbps), making it impractical for many applications.*

*Speech coding is the process of reducing the data rate of digital voice to manageable levels. Parametric speech coders or vocoders utilise a-priori information about the mechanism by which speech is produced in order to achieve extremely efficient compression of speech signals (as low as 1 kbps).*

*The greater part of this thesis comprises an investigation into parametric speech coding. This consisted of a review of the mathematical and heuristic tools used in parametric speech coding, as well as the implementation of an accepted standard algorithm for parametric voice coding.*

*In order to examine avenues of improvement for the existing vocoders, we examined some of the mathematical structure underlying parametric speech coding. Following on from this, we developed a novel approach to parametric speech coding which obtained promising results under both performances of two different encoding algorithms on the two languages objective and subjective evaluation.*

***Index Terms*—Voice Over Internet Protocol (VOIP), LP (Linear Predictor), MELP (Mixed Excitation Linear Predictor).**

### I. INTRODUCTION

voice recognition systems have become increasingly popular as a means of communication between humans and computers. An excellent example of this is the AST automated reservation system developed at the University of Stellenbosch, which makes hotel reservations over the telephone.

It is a well-known problem that the accuracy of these voice recognition systems is adversely affected by the effects of telephone channels.

Therefore it would be advantageous to be able to use digital voice for the recognition system. This could potentially reduce the amount of training data required by reducing the number of telephone channel conditions which must be catered for. At the same time digital transmission of voice could minimize the transmission channel effects, thus improving the clarity of the input voice and improving the overall recognition accuracy of the system.

This need for digital voice communication suggests the implementation of a voice coder suitable for a Voice over Internet Protocol (VOIP) system. Recent changes in Telecommunications legislation have made such systems a highly viable proposition[1].

However, most parametric voice coders have been developed within the context of an Low rate or multi rate implementation to cater for applications where bandwidth is limited.

Multi-language compatibility. Most current voice encoding standards are aimed at European languages or American English. The phonemic richness of the African languages pose a potential challenge and the voice coding should be able to handle this.

## **II. STANDARD VOICE CODING TECHNIQUES**

LPC10e refers to an algorithm which may originally be attributed to Atal and Hanauer [2]. FS1015 and LPC10e have essentially become synonymous

### **Pre emphasis of S**

Speech is pre-emphasised with a first order IIR filter with the following function.

$$H(z) = (1-15/16z^{-1})$$

The purpose of this filter is to improve the numeric stability of the LP analysis. The speech waveform typically exhibits a high-frequency roll-off. Reducing this roll-off decreases the dynamic range of the power spectrum of the input speech, resulting in better modeling of the features in the high frequency regions of the speech spectrum [3].

### **LP Analysis**

The LPC10e standard (FS1015) specifies that a covariance method with synthesis filter stabilization should be used to determine the LP spectrum of the speech. However, most modern implementations instead use an autocorrelation approach due to its improved numerical stability and computational efficiency and since this does not affect the interoperability of the vocoder at all. FS1015 favours a pitch synchronous LP analysis. This means that the position of the LP analysis window is adjusted with respect to the phase of the pitch pulses. This design improves the smoothness of the synthesized speech, since the effect of the glottal excitation spectrum on the LP analysis of the speech is reduced substantially. LPC10e allows pitch ranged between 50 and 400Hz. The pitch estimate is obtained as follows.

1. Low pass filter the speech signal
2. Inverse filter the speech signal with a second order approximation to the optimal 10th order predictor determined by the LP analysis.
3. Calculate the minimum value of the Magnitude Difference Function (MDF)[4]

## **III. FS1016 - CELP**

CELP was first proposed by Atal and Schroeder in their 1985 paper [5]. It uses the same source-filter model as LPC, except that in the case of CELP, the simple buzz-hiss excitation of LPC is replaced by a more sophisticated excitation model.

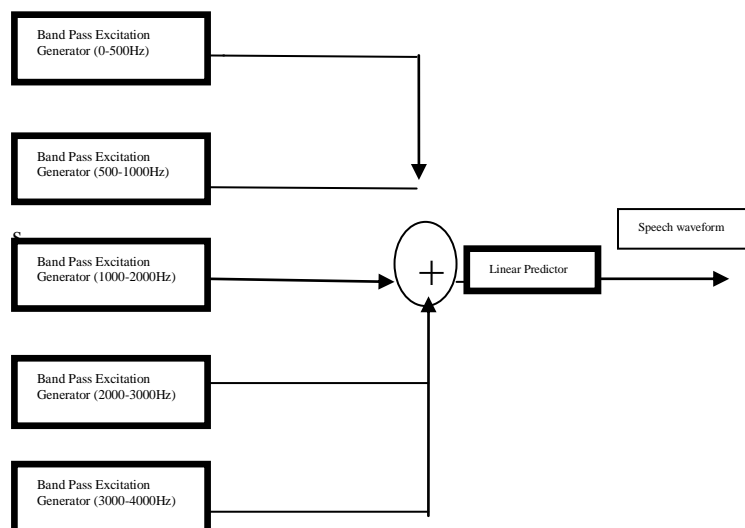
In CELP, the excitation used in each frame is selected by the encoder from a large predetermined codebook of possible excitation sequences. Hence the acronym of Codebook Excitation with Linear Prediction. The typical way in which the excitation codebook entry is chosen is by means of analysis by synthesis.

In traditional open loop analysis methods, an analysis of the speech signal is performed and the excitation sequence is chosen based on the result of this analysis. In the CELP encoder, a more sophisticated closed loop approach is taken. In this approach every possible excitation sequence is passed through the synthesis filter.

#### IV. MELP

The MELP model was originally developed by Alan McCree as a Ph.D project and was published by McCree and Thomas Barnwell in 1995 [6].

After some refinement, it was submitted as a candidate for the new U.S. federal standard at 2.4kbps. MELP officially become a U.S. federal standard in 1997, replacing LPC10e as the standard vocoder to be used in secure and digital voice communication over low bandwidth channels. The draught 2.4kbps MELP standard can be found in [7].



#### Band pass excitation Generator in MELP Synthesis

In the MELP analysis, the input waveform is filtered by a bank of FIR bandpass filters. These filters are identical to the filters used to band-limit the excitation signals. This produces 5 different band-limited approximations of the input speech signal. A voicing strength is determined in each of these band-limited signals. This voicing strength is regarded as the voicing strength for that frequency band.

These band limited excitation waveforms are added together to produce an excitation signal which is partly voiced and partly unvoiced. In this way, the MELP excitation signal is generated as a combination of band pass filtered pulses and band pass filtered white noise. This substantially reduces the harshness of the voicing decision and removes a great deal of the hissiness and buzziness of LPC10e.

In 1998 McCree and DeMartin [8] published an improved MELP vocoder which claimed to produce better speech quality at



1.7kbps. The salient features of this new vocoder are:

### V. IMPROVED PITCH ESTIMATION

A sub-frame based pitch estimation algorithm is used which significantly improves performance in comparison to the pitch tracking used in the Federal Standard. This algorithm minimises the pitch-prediction residual energy over the frame, assuming that the optimal pitch prediction coefficient will be used over every sub-frame lag. This algorithm is substantially more accurate over regions of erratic pitch and speech transitions.

An averaged PSD is used to calculate an estimate of the noise power spectrum. The estimate of the noise PSD is used to design a noise suppression filter. Instead of the 25 bit-per-frame quantisation used in the Federal Standard, a 21bit-per-frame switched predictive quantisation scheme using a theoretically optimized LSF weighting function is used.

### VI. MELP AT 600BPS

In 2001 Chamberlain [9] proposed a 600bps vocoder based on the MELP voice model. In this vocoder, the analysis and synthesis are done on 25ms segments. However, four consecutive speech frames are encoded together in order to exploit the substantial interframe redundancy which may be observed in the MELP speech parameters. A total of 60 bits are used per 100ms encoding super-frame (4 analysis frames). The encoding structure is as follows .

Parameter	No. of bits allocated
Voicing	4
Energy	11
Pitch	7
Spectrum	38

Bit allocation in Chamberlain's 600 bps MELP Vocoder

#### Aperiodic Flag

The aperiodic flag is omitted from this version of MELP. Chamberlain justifies this decision by stating that at this bit-rate, more significant improvements may be obtained by better quantisation of the other speech parameters than by the inclusion of the aperiodic flag.

### V II. BAND-PASS VOICING QUANTISATION

Table shows the probabilities of occurrence of the various band pass voicing states. From the table it is clear that the band-pass voicing may be quantised to only two bits with very little audible distortion. A further gain is achieved by exploiting the inter-frame redundancy of the band-pass voicing parameters. In this way Chamberlain manages to compress  $4 \times 5 = 20$  bandpass voicing bits into only 4 bits. Chamberlain states that at this level of quantisation some audible differences are



heard in the synthesised speech, but that the distortion caused by the band-pass voicing is not offensive.

Voicing Status (Lowest to Highest Band)	Probability of Occurrence
UUUUU	0.15
VUUUU	0.15
VVVUU	0.11
VVVVV	0.41
Other	0.18

(MELP band pass Voicing probability)

### VIII. IMPLEMENTATION OF AN IRREGULAR FRAME RATE VOCODER

In the section we illustrated how we may possibly represent the speech signal accurately with fewer sampling points using irregular sampling of the parameter trajectory.

In this topic we will apply these ideas to the MELP speech production model in order to develop a variable frame-rate vocoder. The development of such a Vocoder requires the following.

1. An algorithm to determine an accurate representation of the feature vector trajectory, by sampling  $p(t)$  at a high sampling rate.
2. A reconstruction algorithm, which can approximate  $p(t)$  from a set of feature trajectory samples,  $\{p[t_1], p[t_2], \dots, p[t_N]\}$ .

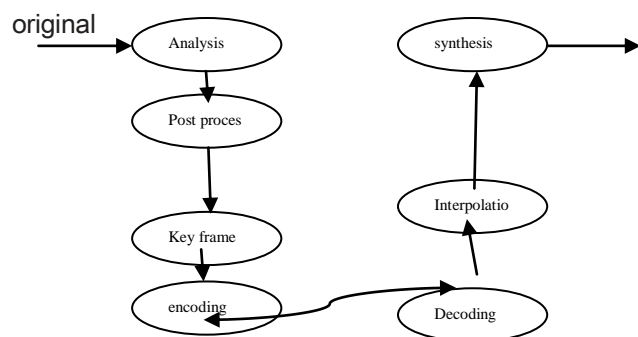
We will refer to this approximation as  $p(t)$

A corresponding decomposition algorithm to determine an optimal set of sampling points  $(t_1, t_2, \dots, t_N)$  so that the reconstruction will be as close as possible to the original for a given frame rate. In contrast to the analysis-by-synthesis approach taken in [10] and [11], we will attempt to determine the sampling points directly from analysis of the feature trajectory. We will refer to the above optimal set of points as the key frames for the speech segment.

This is illustrated in figure.

The way in which this has been implemented is as follows:

1. We adapted the analysis engine of the standard MELP vocoder to determine an over-sampled representation of the parameter trajectory.
2. We used simple linear interpolation to calculate  $p$  from  $\{p[\tau_1], p[\tau_2], \dots, p[\tau_N]\}$ .



**(IS-MELP BLOCK DIAGRAM)**

In the IS-MELP analysis step, the input speech waveform is analysed using the standard MELP analysis. However, the IS-MELP analysis window is advanced by only 2.25 ms (or 18 samples) at a time instead of the 22.5ms (180 samples) by which the standard MELP analysis window is advanced. This results in a tenfold oversampling of the parameter trajectory.

The primary purpose of this over-sampling is that the oversampling allows for more accurate identification of the significant points in the speech parameter trajectory. We determine the feature trajectory in our algorithm by performing MELP analysis on overlapping frames of the speech waveform. The standard MELP analysis is performed on analysis frame of 22.5 ms for every analysis. In our algorithm we attain a high resolution view of the trajectory by advancing the analysis frame by only 2.25ms. This of course leads to substantial redundancy in the feature vector trajectory, analogous to the redundancy produced by over sampling a band limited signal. In order to utilize this redundancy to obtain a more accurate estimation of the trajectory, we will perform a filtering step on the feature trajectory.

**IX. CONCLUSION**

The bit-rate indicate that it is possible to achieve continuous variation of the bit rate and quality of the voice coding system by varying the allowable distortion. Furthermore, this decision may be continuously adjusted at the transmitter without introducing the necessity of transmitting additional information to maintain synchronisation with the receiver.

The most significant disadvantage of the IS-MELP vocoder is the difficulty of relating the distortion thresholds to a fixed bit-rate. Since there is no simple mathematical function which determines the bit-rate from a set of thresholds, the bit rate produced by a threshold set must be evaluated empirically. However, in an application environment, this problem could be circumvented in one of two ways:

1. By adaptively altering the thresholds in order to produce the desired bit-rate.
2. By storing optimised threshold sets for various bit-rates and loading an appropriate threshold set for the desired bit-rate.

While the IS-MELP algorithm has produced results comparable to those of the regular MELP algorithm, and in some cases demonstrated superior performance, the performance, particularly at low frames rates, was found to be unsatisfactory. This was most apparent from the subjective tests. We feel that substantial improvement of the IS-MELP algorithm may still be achieved.

**REFERENCE**

1. Government, S. A., Policy announcement by the minister of Communications, Drivy Matsepe -Casaburri. <http://www.info.gov.za/speeches/2004/04090310151004.htm>, September 2004
2. ATAL, B. S. and HANAUER, S. L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." Journal of the Accoustic Society of America, 1972.
3. CHU, W. C., Speech Coding Algorithms. Hoboken: Wiley, 2003.
4. ATAL, B. and SCHROEDER, M., "Predictive Coding of Speech Signals." Report of the 6th International Conference on Accoustics, 1968.
5. ATAL,B, "Efficient Coding Of LPC parameters by Temporal Decomposition." IEEE ICASSP, 1985.
6. MCCREE, A. and III, T. P. B., "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding." IEEE Transactions on Speech and Audio Processing, July 1995. 7.Publication , F I P S, "Analog to Digital Conversion of voice by 2400 bit/s MELP" June 1997.
8. MCCREE, A. and MARTIN, J. C. D., "A 1.7 kB/s MELP Coder with improved Aalysis and Quantisation." IEEE ICASSP, 1998.
9. CHAMBERLAIN, M., "A 600 bps MELP vocoder for use on HF channels." IEEE Military Communications Conference, October 2001, Vol. 1.
10. ATAL, B., "Efficient Coding of LPC Parameters by Temporal Decomposition." IEEE ICASSP, 1985.
11. CHENG, Y.-M. and O'SHAUGHNESSY, D., "On 450-600b/s Natural Sounding Speech Coding." IEEE Trans. Speech Audio Processing, April 1993.