Significance of Vocoders in Mobile Communication

R. Chinna Rao, D. Elizabeth Rani, S. Srinivasa Rao

Abstract*: Signal Processing finds its applications in many fields of engineering like Communications, Multimedia processing like audio, speech, video compression. It is the constant endeavor in compression techniques to retain highest quality of media through lowest bits so that communication band width utilization can be maximized. The most basic form of end user communication in the mobile industry has been voice calls. Speech signals captured at the microphone are sampled at 8 or 16 KHz and using 16 bits per sample presents 128 KHz of information. Transmitting all of it is very inefficient usage of the communication channel. There are multiple ways to encode the speech input at lower bit rates using Source Coding and Waveform coding techniques. This paper is scoped to practically simulate some of the basic and advanced signal processing concepts and apply them to the speech signal processing domain to minimize the vocoders packet exchange bit rate. Existing techniques are studied and a new schema is proposed which reduce the vocoders packet exchange bit rate further.

Keywords*: Formant, Larynx, Vocal tract, Phonology, Velum, Liner predictive coding.

I. INTRODUCTION

A. Vocal Organs

A simplified cross sectional view of the human vocal system is shown in figure 1. The major vocal organs and their function are described in brief below:

Lungs: Source of air/excitation

Larynx: which are the vibrating parts of the vocal tract and they flap together as air is forced through the slit between the vocal cords, called the glottis.

Oral Tract: This is an acoustic tube of varying cross sectional area, a path comprising the jaws, tongue and lips that air from lungs takes to produce sound through the lips.

Nasal Tract : This is non-uniform acoustic path of fixed cross sectional area that the air from lungs can take to produce sound. This is the acoustic path that ends at the nose.

Velum: This is a small movable flap of skin which controls the acoustic coupling between the oral and nasal tract.

B. Mechanism of Speech Production

The lungs pump in air through the rest of the wind pipe. This air flow is the source of energy for sound generation.. The rest of the oral cavity comprising of the wind pipe, jaws, tongue and lips together form a hollow pipe of varying cross-sectional area which converts the energy coming through the glottis into voice sound. Alternately, the vocal cords do not vibrate and simply pass the air through the glottis. At the velum, part of the air passes through the oral cavity and part of it through the nasal cavity. The sounds produced as a combination of these different structures can be classified as voiced, unvoiced, plosive etc.

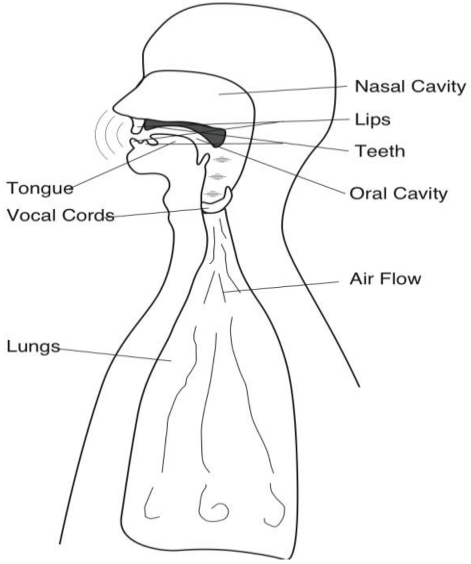


Fig.1. Human Vocal System

In the production of non-nasal sounds, the velum seals off the nasal tract and there is single sound transmission path via the lips. For nasalized sounds, the sound comes from both the lips and nose. According to mode of excitation, speech can be broadly classified as voiced and unvoiced sounds.

C. Voiced Sound

In production of voiced sounds like ‘aah’, ‘oh’, the vocal cords are made to vibrate opening and closing the glottis in a regular fashion. This vibration of the vocal cords gives rise to a fundamental frequency and a set of harmonics called the pitch period. Pitch range in males is from 50Hz to 250 Hz, in females it goes as high as 500Hz.

D. Unvoiced Sounds

In this case, the vocal cords do not vibrate as air is pushed through the glottis in a continuous form. There is no vibration and the wind pipe does not present any resonance frequency and only the constriction of the lips, produces the sound.

Revised Manuscript Received on December 13, 2019.

***** Correspondence Author

R. Chinna Rao*, , Assistant Professor in Department of ECE at MRCET, Secunderabad, Telangana, India.

Dr. D. Elizabeth Rani, Professor, Department of EIE, GITAM University, Visakhapatnam, Andhrapradesh, India.

Dr. S. Srinivasa Rao, Professor, Department of ECE, MRCET, Secunderabad, Telangana, India.

E. Source-filter model of speech production

The human speech production mechanism can be modeled as a simple source-filter structure, where to produce sounds.

Excitation generator: The vocal cords vibration in case of voiced sounds and the air passing through the open glottis in case of unvoiced sound can be modeled simply as the excitation generator which pushes either periodic pulses or random noise through the vocal tract.

Vocal Tract*:* The vocal tract(oral cavity or nasal tract) can be thought of as an open pipe of either varying or constant cross sectional area which has resonances called formants.

II. PARAMETRIC SPEECH ANALYSIS

Given the physiological constraints of of speech generation, the movement of the jaws, tongue and lips, the speech signal can be considered stationary over a significantly small interval of time (10-30 ms). In cellular networks, the standard protocols dictate packet exchanges to happen every 20 miliseconds. Thus, for speech analysis, short 20 milisecond frames are used for analysis and encoding. Speech signal can be analyzed in a time domain as well as in the frequency domain.

A. Time Domain Analysis

While frequency domain processing using bank of filters, DFT, and cepstral processing are popular, time domain measures such as zero crossing count, signal energy, autocorrelation can be used to extract limited but useful information about the speech signal. The powerful technique of LP, it is essentially an efficient time domain waveform coding technique and is used to estimate the frequency response of the vocal tract.

B. Zero Crossing Count

The time domain analysis of short frames of speech signal can be used extract certain features like Zero Crossing Count. It can be used to get some idea of whether the signal source is part of voiced or unvoiced sound, because the voiced sound will be a more rapidly changing signal than an unvoiced speech segment. Example of zero crossing count: Zero Crossing Count has a good value of 12 in the first zoomed speech segment shown on the left bottom whereas in silence zone it is almost zero as shown in the right bottom figure. In other segments with more speech content, the ZCC went upto 21 within a 20 milisecond speech frame.

III. LINEAR PREDICTION ANALYSIS & RESULT

Linear Prediction is a very powerful technique which is used in almost every speech coding (compression) algorithm. It is used to identify a small set of coefficients with which future samples of speech can be predicted based on the past samples and by applying these coefficients so that the mse between the actual sample and the estimate sample is minimized. The problem boils down to finding the

A.Number of past samples to use for prediction. In other words, we need to narrow on the number of coefficients that need to be used to minimize the energy of the error.

B.The find the coefficients.

In case of speech encoding these coefficients are transformed to another domain, example, and reflection coefficients and transmitted over the communication channel so that the other end can reconstruct the signal back from minimal number of information bits. Based primitively on the LPC coefficients, we have the basic LPC10 vocoder whose block diagram is shown below.

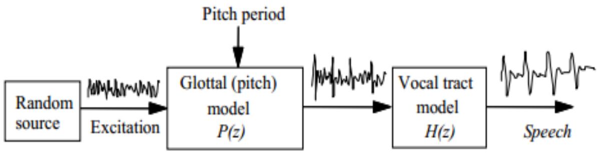


Fig. 2. Source Filter Model of Speech Generation

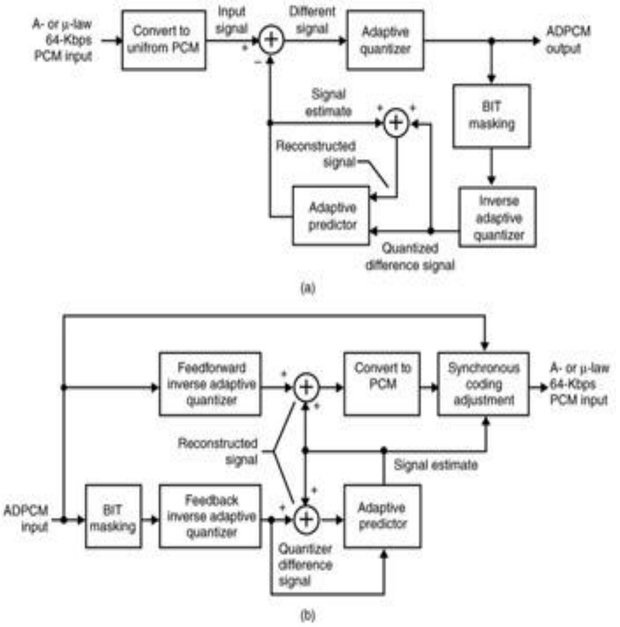

A- or µ-law 64-Kbps PCM input
Convert to unifrom PCM
Input signal
Different signal
Adaptive quantizer
ADPCM output
BIT masking
Signal estimate
Reconstructed signal
Adaptive predictor
Inverse adaptive quantizer
Quantized difference signal
(a)
Feedforward inverse adaptive quantizer
Convert to PCM
Synchronous coding adjustment
A- or µ-law 64-Kbps PCM input
Reconstructed signal
Signal estimate
ADPCM input
BIT masking
Feedback inverse adaptive quantizer
Adaptive predictor
Quantizer difference signal
(b)

Fig.3. LPC10Encoder and Decoder

A. LPC 10 Coder

Block diagram of LPC 10 Encoder and Decoder is shown in above figure 2. The output quality of LPC 10 sounds unnatural because normal voice is neither completely voiced nor completely unvoiced, yet the above model takes a completely binary decision. The LPC-10 is able to achieve lower bit rates but at the expense of quality. Ideally, one would like to achieve both low bit-rates and good quality, which is what, has been the focus of speech coding research over the years.

B. Speech Encoding/Decoding CELP structure

Given the mature nature of the field of Speech Processing, immense focus has already gone into minimizing the bit rate of vocoders while maintaining minimal quality degradation. The CELP structure and other coding standards studied focus on different ways to codify the parameters of each speech frame using optimal number of bits. The codebook tables were developed so that excitations could be codified more effectively; the LPC parameters are transformed as reflection coefficients to minimize impact of quantization sensitivity of LPC parameters. These schemes are focused on the specific speech frame under analysis and the techniques are for compression lead to an ‘intra-frame’ compression. The purpose of speech encoding is to codify a speech segment in the minimal number of bits possible so that the receiver can reconstruct the speech segment with highest voice quality. In order to achieve this CELP structure is built around the human speech production model. Based on the CELP structure the transmitter identifies and codes

1. The LPC coefficients
2. The correct excitation that should be passed through the filter with the LPC coefficients as filter coefficients. This is generally a pseudo random noise sequence selected from a code book. The LPC coefficients form part of the short-term predictor.
3. The codebook index for the pitch information of the speech signal in case of voiced segments. This is called the long-term predictor. This information is fundamental to the speech analysis by synthesis model, which is widely deployed in many of the speech encoders.

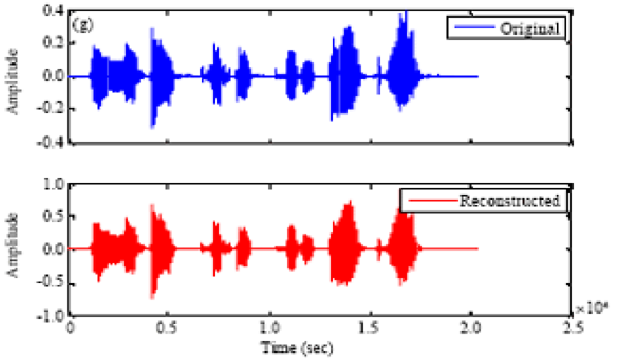
C. Linear Prediction Coefficients

The linear prediction technique as described in section 3.2 reduce the short-term correlation in the transmitted signal but to achieve lower bit rate, the long term correlation should also be reduced. Given the sensitive nature of LPC coefficients to quantization errors, they are transformed to LSF pairs or to reflection coefficients and made part of the encoded packet.

D. Pitch

After subtracting the Linear Predicted Signal from the original signal, the residual signal is found to have quasi-periodic nature as it is related to the excitation from the lungs. Because of the vibration we can see a ‘periodic’ nature of the residual signal. The residual signal is evaluated for calculating the periodicity known as the pitch period. This is done by performing autocorrelation on the short term residual signal as shown below.

The index which gives the maximum value for auto-correlation is chosen as the Long Term Predictor value. The gain can be obtained using the standard optimal filtering techniques. While the LPC coefficients are on a per 20 ms frame basis, the LTP are on every 5ms sub frame basis.



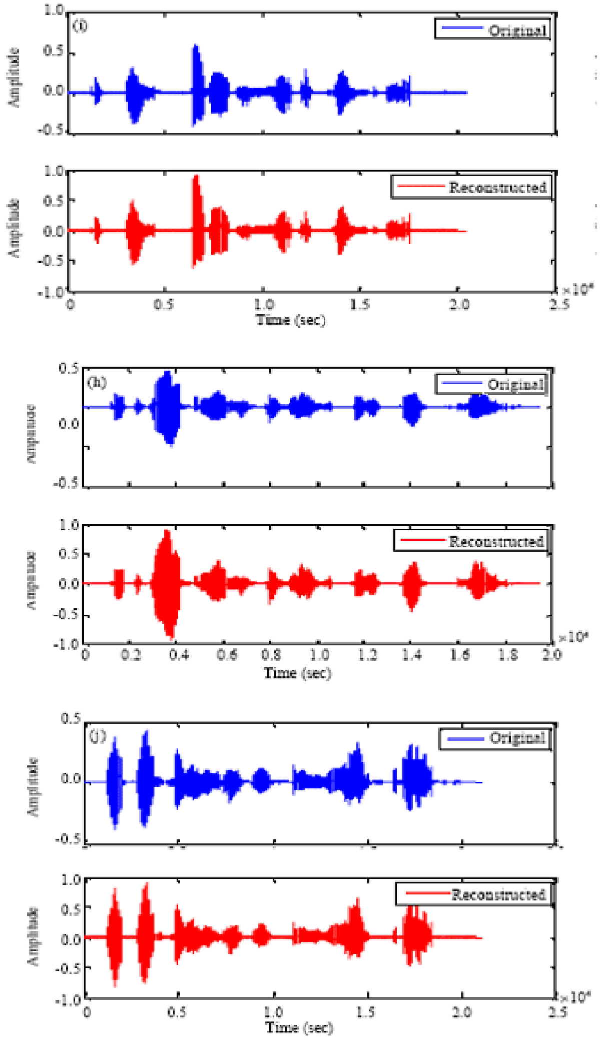


Fig.4. Speech Signal Analysis

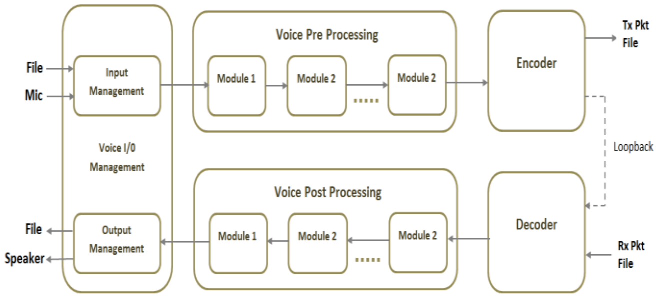


Fig.5. Basic Frameworks of Vocoders

Table I: Analysis of Speech Coders

Codec	Standard body/year	Type	BW	Bit rate (kb/s)	Bits per sample /frame
G.722	ITU/1988	SB-ADPCM	WB	64	8
GSM	ETS/2000	ACELP	NB	12.2	244
ILBC	IETF/2004	CELP	NB	13.33	400
SPEEX		CELP	WB	2-44	20
G.711 U AW	ITU/1972	PCM	NB	64	8
G.711 ALAW	ITU/2008	PCM	WB	64	320

Retrieval Number B10181292S319/2019©BEIESP
DOI: 10.35940/ijitee.B1018.1292S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

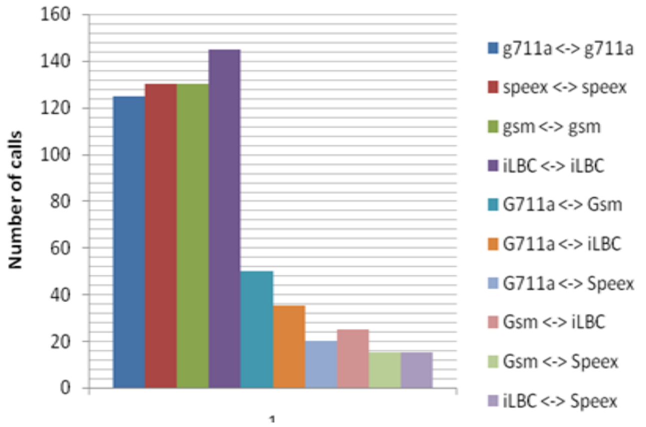


Fig.6. Bar graph showing Analysis of Speech Coders

E. Excitation

What remains after removing the short term and long term predictors is the residual signal. This residual signal is what acts as the excitation for the decoder. The decoder has to take this excitation signal, apply the LTP and pass the result through the LPC filter to get the speech signal.

In order to transmit the residual signal, tables of 40 samples(5ms) are developed which act as sources of excitation. The encoder identifies the most closely matching excitation and codifies its index in the packet.

The decoder maintains a local replica of the codebook table and using the excitation from the table and the LTP and LPC parameters in the packet, generates the speech signal. Because the excitation at the decoder is picked using a 'code', this is called Code Excited Linear Prediction (CELP) technique. Techniques like CELP focus on intra-frame compression.

Other techniques like DTX focus on inter-frame compression but this gets kicked in only during silence. This is to take advantage of the fact that during voice communications it is only one person speaking at a given point of time, so the packet transmission from the listener's end can be optimized. The idea employed here is to first classify a frame of samples as containing active speech or as a silence frame using various VAD techniques. In case the segment is identified as a silence frame, there is no need to perform complex LPC computations. Under DTX scheme, in case of silence frame, a full-fledged vocoder packet is not transmitted, instead only an indication is sent to suggest the ONSET or CONTINUATION of silence period. The decision of DTX schemes kicking in is based on the VAD parameter of the speech frame under consideration.

IV. DISCONTINUOUS TRANSMISSION

The concept of Discontinuous Transmission (DTX) is an efficient way of achieving spectral efficiency where bandwidth is expensive. The DTX is ON only active packets Remaining is needed not to transmit. The Voice Activity Detector (VAD) is the most important part of the current DTX schemes. The design of the activity detector has to balance the risk of clipping speech segment (wrongly identifying speech as noise) against the risk that noise is classified as speech. When the VAD is used to switch the transmitter on and off, the effect will be a step change for the background noise at the receiver. The sudden change in noise level can be perceived as annoying by the receiver. A way to mask this is to generate some noise at the decoder when the transmitter is turned off. However this noise should be similar in nature to the background noise at the transmitter side. Therefore, when silence is found, the Tx should periodically transmit the average for accurate reconstruction of the background noise. This noise is called comfort noise.

V. PACKET LOSS CONCEALMENT

Packet Loss Concealment (PLC) Algorithms or Frame Erasure Concealment Algorithms generate synthetic speech signal to cover missing data (packet drops due to network characteristics and conditions). Since speech signals are locally stationary, it is possible to use the signal's past history to generate a reasonable approximation to the missing segment, without audible artifacts, unless the erasure falls in the region of rapidly changing speech signal.

VI. CONCLUSION

In this paper simulate and studied different types of speech analysis like parametric speech analysis, Time domain speech analysis and linear predictive analysis of LPC and CELP and also using DTX techniques. And also simulate different types of Vocoders performances using Wireshark packet analyzer with Asterisk PBX.

REFERENCES

1. Usman, M., Zubair, M., Shiblee, M., Rodrigues, P., Jaffar, S."Probabilistic modeling of speech in spectral domain using maximum likelihood estimation", Symmetry, 10(12) Volume 10, Issue 12 , pp.1-15, **2018**.
2. J. P. Campbell, Jr., "Speaker Recognition: A Tutorial", Proceedings of The IEEE, Vol.85, No.9, pp.1437-1462, Sept.1997.
3. Koji Kitayama, Masataka Goto, Katunobu Itou and Tetsunori Kobayashi,"Speech Starter: Noise-Robust Endpoint Detection by Using Filled Pauses", Eurospeech 2003, Geneva, pp. 1237-1240.
4. S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition", in Proc. ICASSP2002, vol. 4, 2002, pp. 3808–3811.
5. Martin, D. Charlet, and L. Mauuary, "Robust speech / non-speech detection using LDA applied to MFCC", in Proc. ICASSP2001}, vol. 1, 2001, pp. 237–240.
6. K. Ishizaka and J.L Flanagan, "Synthesis of voiced Sounds from a Two-Mass Model of the Vocal Chords," Bell System Technical J., 50(6): 1233-1268, July-Aug., 1972.
7. Gopatoti, A., Ramadass, N. "Performance of adaptive subband thresholding technique in image denoising", Journal of Advanced Research in Dynamical and Control Systems, Vol. 9, No. 12, (2017), pp.151-157.
8. Atal, B.; Rabiner, L., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition" Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume: 24 , Issue: 3 , Jun 1976, Pages: 201 - 212.
9. Billa, P., Gopatoti, A. "3D MR images denoising using adaptive blockwise approached non-local means (ABNLM) filter for spatially varying noise levels", International Journal of Innovative Technology and Exploring Engineering, Vol.8 No. 4(2019), pp.112-118.
10. Raviteja, M.L., Nagaraju, N., Gopatoti, A., Paparao, N.," Processing algorithm and data acquisition for laser range sensor", International Journal of Innovative Technology and Exploring Engineering, Vol.8 No. 3(2019), pp.53-57.

AUTHORS PROFILE

Mr.R. Chinna Rao, received his B.Tech degree in Electronics & communication from JNT University. M.Tech from Malla Reddy College of Engineering & Technology. He is currently working as Assistant professor, Dept. of ECE in Malla reddy College of Engineering and Technology, Secunderabad, India

Dr.Elizabath Rani , received the B.Tech degree from Madurai Kamaraj University,M.Tech from Bharathiar University and Ph.D from Andhra University Visakhapatnam. Presently working as Professor and Head of the Department at Gandhi Institute of Technology and Management, Visakhapatnam. She has 31 years of experience in the field of teaching. She is a member of professional bodies like MISTE, IETE and SEMCE(I).

Dr. S.Srinivasa Rao, received the B.Tech degree from Madras Institute of Technology, Anna University, and the M.Tech and Ph.D from JNTU Hyderabad, Telangana, India. Presently working as Professor and Head of the Department at Malla Reddy College of Engineering and Technology, Secunderabad. He has 24 years of experience in the field of teaching. He is a member of professional bodies like IEEE, ISTE and IETE.