A low bit rate speech coding and its synthesis based on Hidden Markov Models

Firos A
Department of Computer Science & Engineering,
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India

Prof. Utpal Bhattacharjee
Department of Computer Science & Engineering,
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India

***Abstract*—**This paper presents a approach for developing semantic based speech coding technique by preserving its prosodic features. GMM model will be incorporated to identify the semantic features and prosody of the windowed speech. LPC analysis will be done parallel. Fussy k–mean clustering will be done in the extracted features. ANN will be utilized to identify the best features for encoding. Using such semantic based coding will considerably reduce the processing requirements while encoding.

Keywords— speech coding; semantics; fuzzy k-mean clusturing; Windowing; ANN

I. INTRODUCTION

In recent years, the advancement in communication is being driving the evolution of networking technologies. Voice communications is an indispensable part of network communication. The competitive section of various compression methods confirms that the majority still adheres to traditional art practice, related parameters and data compression on speech signals. Speech to text conversion system is extensively used in many applications. But its strength is not widely explored for speech compression techniques. Even though the performance issues of the semantics based speech compression is a matter of concern, this type of coding methods will be highly anticipated for low bandwidth networks. Emergence of mobile communication technology has posed new challenges for developing better bit rate efficient coding standards. The major challenge in usual speech coding standard is its performance degradation over time. This major challenge can be addressed by introducing semantic based speech analysis while encoding.

In realistic environments, there exist more complicated application scenarios, such as bandwidth degrading communication networks, which may lead to poor quality to voice communication. This paper proposes scalable semantic based speech coding technique wherein, the decoding will be done with the help of fuzzy system and ANN. The objective evaluation of the results shows that, the proposed technique provides high robustness against packet loss and also achieves a better performance while doing voice communication in poor bandwidth.

This paper is organized as follows. Section II describes various existing speech coding techniques. Section III describes the proposed encoding and decoding algorithms. Finally, Section IV concludes the paper.

II. EXISTING SPEECH CODING TECHNIQUES

Due to biological constraints, a compression scheme is required to adjust the wide dynamic range (DR) of input signals to a desirable level [1] .Real-time speech coding for VoIP is no mean milestone for an modern speech coding technique. Particularly when almost all speech coders many speech coders use a model of a vocal tract, and chose parameters of the model to best match the signal. In the recent years, speech coders categorized the coding method to falls under two categories namely narrow band and wide band speech coding. So there was an array of established names of coding like G.723.1, G.726, G.728, G.729, iLB etc. though it lacked the expectation of high definition video conferencing systems.

Speech compression came by way of some fine works. Like the Internet Low Bitrate Codec (iLBC) with its handling of lost frames through graceful speech quality degradation. When it was introduced in the year 2004, people thought “Is this coding waiting to create speech’s essence by scripting its existence?”. Lost frames often occur in connection with lost or delayed IP packets. Ordinary low-bitrate codecs exploit dependencies between speech frames, which cause errors to propagate when packets are lost or delayed. In contrast, iLBC-encoded speech frames are independent and so this problem will not occur.. But the disadvantage of iLBc is that it does not have the adoption rate .As unsettling as the adoption rate in iLBc, we have its competitor, G.729. As though G.729 been assembled as a spiteful joke of Speech quality decrease by marginally, it is been well accepted

G.711 is a Pulse code modulation (PCM) of voice frequencies on a 64 kbps channel. G.711 uses a sampling rate of 8,000 samples per second. Non-uniform quantization with 8 bits is used to represent each sample, resulting in a 64 kbit/s bit rate. There are two types of standard compression algorithms are used here. (1) μ-law algorithm (2) A-law algorithm. What takes attention here is that G.11 is designed to deliver precise transmission of speech. At implementation sideG.711 frayed Very low processing overheads, although it give poor network efficiency as grave as it's edgy. However, G.711.1 is an extension version of G.711 is more taken up with addition of narrowband and/or wideband (16000 samples/s) enhancements, which leading to data rates of 64, 80

or 96 kbit/s. G.711.1 's embedded bit stream structured in three layers corresponding to three available bit rates: 64, 80 and 96 kbit/s with its adaptive bit allocation for enhancing the quality of the base layer in the lower band and weighted vector quantization coding of the higher band based on modified discrete cosine transformation (MDCT) that run and turn sharply are striking, too.

G.711.1 was as dynamic and unexpected as ever but G.722, with just a subtle deviation - Technology of the codec is based on split band ADPCM- turns coding's robustness head into a remarkable work. G.722 is delightfully, It is useful in fixed network voice over IP applications, where the required bandwidth is typically not prohibitive, while it offers a significant improvement in speech quality over older narrowband codecs such as G.711. G.722.1is an extension version of G.722 is a transform-based compressor that is optimized for both speech and music. The computational complexity is quite low and the algorithmic delay end-to-end is 40 ms. G.722.2 is also referred as AMR-WB. It is a speech coding standard developed after the AMR using same technology like ACELP. G.722.1 and G.722.2 stand out in their uncluttered arena of speech coding.

Compression of a high order comes from G.723.1. There's technical finesse in G.723.1, while much effective in the audio portion of videoconferencing/telephony over public telephone (POTS).. but it suffers lower quality than many other codecs at similar data rates. G.723.1 seems to float across in layers, overlapping, coalescing, drifting apart and fading.

G.723.1 was also as dynamic and unexpected as ever but G.726, with voice at rates of 16, 24, 32, and 40 kbit/s.G.721 and G.723 had been folded into G.726 turns coding's robustness even better. G.723.1 is delightfully, 32 Kbits which is half the rate of G.711 codec and hence increasing the usable network capacity by 100%, while it offers a significant improvement in speech quality over older narrowband codecs such as G.723.1.. The computational complexity here also is quite low and the algorithmic delay end-to-end around 40 ms.

The competitive section of various compression methods confirms that the majority still adhere to traditional art practice, related parameters and data compression on speech signals, perhaps because of the categories laid down by the gallery of traditional compression methods. Another interesting trend is of Adaptive Dynamic Range Compression. Here, dynamic nonlinear time-variant operator, such as a dynamic range compressor, can be inverted using an explicit signal model. By knowing the model parameters that were used for compression one is able to recover the original uncompressed signal from a "broadcast" signal with high numerical accuracy and very low computational complexity [2].

But the winners in compression mechanism in this study is , G.729 and iLBc, proves how real time speech coding can be in his understated yet sinister allegory about the theory of speech coding. G.729 and iLBc uses low delay for compression of speech data as low as 10 milliseconds. Because of its lower bandwidth around 8 kbps it mostly used in Voice over IP (VoIP) applications for its low bandwidth requirement.

III. PROPOSED ALGORITHM

.The proposed algorithm consists of two modules. One is Encoding module, in which input speech is processed and the resultant compressed speech output. Other one is a standard decoding method to map the encoded speech to synthesized speech for real time voice communication. The proposed algorithm is summarized in Fig. 1.

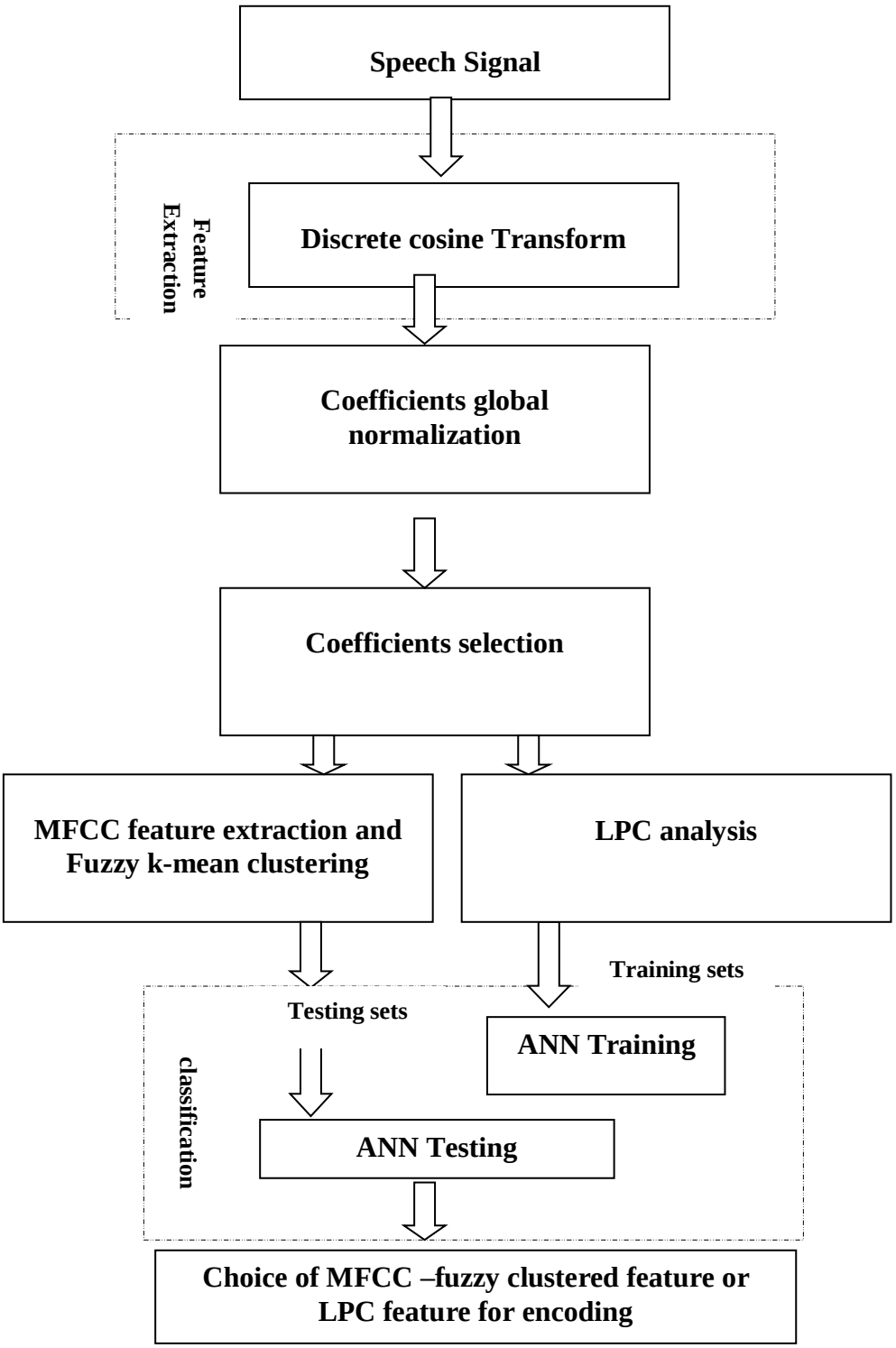
Step 1: as we know with LPC $e(n) = x(n) - \sum_{k=1}^{p} \alpha_k \; x(n-k)$, When $\alpha_k$reduced e(n) reduced. The smaller the error , we have better set of predictors $\alpha_1, \alpha_2, \alpha_3$,----------$\alpha_p$ are slowly varying according to the syllable. These parameters are to be found with the help of LPC .in z-transform X(z) =$\frac{1}{1-\sum_{k=1}^{p} \alpha_k z^{-k}} \; E(z)$, we gets the signal in frequency domain. So, ultimately we will get We have p equations and p unknowns ($\alpha_1, \alpha_2, \alpha_3$,---------- α_p) Every 20ms we have to find $\alpha_k' s$, which is not computationally easy. So we will go for auto correlation method. $S_n\ (m)$=S(m+n)w(m) ; where (m) is the window ; 0<=m<=N-1. So we have E_n=$\sum_{m=0}^{N+p-1} S_n^2(m)$. With this we will get $\emptyset_n(i,k)$= $R_n(i-k)$, where R_n (k)= ∑_(m=0)^(N-1-k)⁡〖 S_n (m) 〗 S_n (m+k) ; R_n (k) is going to be even function. With this for i—1,2,----p we will get a matrix

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \dots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \dots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \dots & R_n(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & & R_n(0) \end{bmatrix} \mathbf{X} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ . \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ . \\ R_n(p) \end{bmatrix}$$

interestingly , since the diagional eements are the some and its a Toeplits matrix , its computationally easy for LPC for computing its α_s

Step 2: The purpose of speech emotion recognition system is to automatically classify speaker's utterances into five emotional states such as disgust, boredom, sadness, neutral, and happiness [4] . The basic assumption is that there is a set of objectively measurable voice parameters that reflect the affective state a person is currently experiencing [5]. At the time LPC is going ahead with the analysis of probabilistic features Gaussian Mixture Model (GMM)-based emotional voice analysis will be done in the same frame to find the prosodic features. Simultaneously the features of the speech signal are extracted by the MFCC block. The total number of samples chosen in a frame is 256 and overlapping samples with the adjacent frame will be 128. We acquire MFCC cepstral coefficients at the output of MFCC block. In GMM, K-mean algorithm is used to obtain a cluster number specific to each observation vector and sets the centriod of the observation vector. After clustering the model, it returns one centroid for each of the cluster K and refers to the cluster number closest to it. K-mean algorithm is described as the squared distances between each observation vector and its centroids. In the training section parameters of GMM model are produced iteratively by expectation-maximization (EM) algorithm. Euclidean distance is found out between

observation vector and its cluster centriods to match the spoken word with the present database[3]. The word which is recognized is taken into another matrix $\begin{matrix} w_1 \\ w_2 \\ w_3 \\ . \\ w_p \end{matrix}$, and its corresponding emotion is $\begin{matrix} e_1 \\ e_2 \\ e_3 \\ . \\ e_p \end{matrix}$



Honey bee The Encoding process

Step 3: the matrices α ,e and w will be taken into feed forward neural network, A feed forward neural network algorithm includes the following steps:

1. Initialize weights and biases to small random numbers.

2. Present a training data to neural network and calculate the output by propagating the input forward.

3. changing in numbers of hidden layers and transfer function for every hidden layer and for output layer and also changing in number of neurons in every hidden layer until reach to maximum recognition and language identification rate or to minimum error.

ANN used here will advice to select the set {w,e} or α to be considered as the encoded signal for propagation.

The decoding module of the proposed algorithm woks as flows.

Step 1: Analysis will be done to on the received speech to identify the decoding is to be done with synthesis of {w,e} or α

Step 2: if step 1 says the frame is synthesized with α , usual LPC synthesis will be done . otherwise {w,e} will be sythesissd with the help of GMM

Information from speech recognition can be used in various ways in state-of-the-art speaker recognition systems [6]. The emotion recognition system must be capable of recognizing at least six basic emotions (happiness, anger, surprise, disgust, fear, sadness) and the neutral circumstances [7].

IV. CONCLUSION

In this paper, a novel semantic based speech compression methods is presented for real-time voice compression which can be used in VoIP. The study considered the strength of both fuzzy k-mean clustering as well as the feed forward artificial neural network, in encoding as well as the decoding of the speech signal.

The quality of speech in the communication system is a matter of concern when it deals with systems like low bandwidth virtual computing infrastructure. For voice communication system encoding may consume comparatively more load than decoding. So, this study deals with processing efficiency at encoding stage also, with the help of artificial neural networks.

This paper proposes a mechanism where the encoding of the speech will be done with the help of its semantic and emotional content, which in turn will help in synthesizing a better quality voice output. This can be accomplished using GMM where in the semantics of the speech may be identified with its emotional content. Since the accuracy is a concern in GMM, LPC also will be incorporated and a better choice of either GMM or LPC feature for decoding will be done with the help of ANN.

The conclusion is that the semantic analysis of the speech content including its emotions parallel with the processing of usual LPC encoding will give better bit rate efficient real-time coding.. Using this approach, it is found that voice communication system can provide better performance especially when the decoding is highly computational and frequent in communication.

REFERENCES

[1] Ying-Hui Lai, Fei Chen , Yu Tsao, ``Adaptive Dynamic Range Compression for Improving Envelope-Based Speech Perception: Implications for Cochlear Implants," Springer, Emerging Technology and Architecture for Big-data Analytics, pp. 191-214, April 2017.

[2] Stanislaw Gorlow ; Joshua D. Reiss ."Model-Based Inversion of Dynamic Range Compression" IEEE, IEEE Transactions on Audio, Speech, and Language Processing , Page(s): 1434 - 1444 ,Volume: 21 Issue: 7, July 2013.

[3] Virendra Chauhan, Shobhana Dwivedi, Pooja Karale, Prof. S.M. Potdar "SPEECH TO TEXT CONVERTER USING GAUSSIAN MIXTURE MODEL(GMM) ", International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, Vol. 2, Issue 3, May-Jun 2012, pp.1169-1173.

[4] Peipei Shen, Zhou Changjun, Xiong Chen," Automatic Speech Emotion Recognition Using Support Vector Machine" IEEE International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT) volume2 , Page(s) : 621 - 625 , 12-14 Aug. 2011.

[5] Akalpita Das, Purnendu Acharjee , Laba Kr. Thakuria , " A brief study on speech emotion recognition" , International Journal of Scientific and Engineering Research(IJSER), Volume 5, Issue 1,pg-339-343, January-2014.

[6] Kshamamayee Dash, Debananda Padhi , Bhoomika Panda, Prof. Sanghamitra Mohanty, " Speaker Identification using Mel Frequency Cepstral Coefficient and BPNN", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, pg.-326-332, April 2012.

[7] Vinay, Shilpi Gupta, Anu Mehra,"Gender Specific Emotion Recognition Through Speech Signals", IEEE International Conference on Signal Processing and Integrated Networks (SPIN), 2014 , Page(s):727 – 733, 20-21 Feb. 2014.

[8] Norhaslinda Kamaruddin, Abdul wahab Rahman,Nor Sakinah Abdullah,"Speech emotion identification analysis based on different spectral feature extraction methods", IEEE Information and Communication Technology for The Muslim World, 2014 The 5th International Conference, Pages:1-5, 2014.

[9] A. D. Dileep, C. Chandra Sekhar, "GMM Based Intermediate Matching Kernel for Classification of Varying Length Patterns of Long Duration Speech Using Support Vector Machines", IEEE Transactions on Neural Networks and Learning Systems, Volume: 25, Issue: 8,Pages: 1421 -1432, 2014.

[10] S.Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh "Speech Emotion Recognition" IEEE International Conference on Advances in Electronics, Computers and Communications (ICAECC), Page(s): 1-4, 2014 .

[11] S.Sravan Kumar, T.RangaBabu , "Emotion and Gender Recognition of Speech Signals Using SVM ", International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 4, Issue 3, pg.- 128-137 May 2015.

[12] R.Banse, K.R.Scherer, "Acoustic profiles in vocal emotion expression", Journal of Personality and Social Psychology, Vol.70, 614-636, 1996

[13] T.Bänziger, K.R.Scherer, "The role of intonation in emotional expression", Speech Communication, Vol.46, 252-267, 2005

[14] F.Yu, E.Chang, Y.Xu, H.Shum, "Emotion detection from speech to enrich multimedia content", Lecture Notes In Computer Science,Vol.2195, 550-557, 2001

[15] D.Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", Speech Coding and Synthesis, 1995

[16] S.Kim, P.Georgiou, S.Lee, S.Narayanan. "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece, 2007

[17] L.R.Rabiner and B.H.Juang. "Fundamentals of Speech Recognition", Upper Saddle River; NJ: Prentice-Hall, 1993

[18] V.A Petrushin, "Emotional Recognition in Speech Signal: Experimental Study, Development, and Application", ICSLP-2000, Vol.2,222-225, 2000 J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[19] Unknown," VOICEBOX: Speech Processing Toolbox for MATLAB" , http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[20] Unknown," PLP and RASTA (and MFCC, and inversion) in Matlab using melfcc.m and invmelfcc.m" http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/