Impact of Semantic Coding of Emotional Speech on Speech Coding Performance

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

Abstract -*This paper presents a technique for solving the real time computational difficulty of speech coding standards in semantic level by preserving its prosodic features. LPC analysis will be done to identify the feature of the input speech. The proposal takes the GMM model to identify the semantic features and prosody of the input speech.. ANN will be utilized to identify the best features for encoding. Using such semantic based coding will highly reduce the computational overhead in speech coders.*

Keywords — *Speech coding; G.723.1, iLBC; fuzzy clustering; Windowing; ANN.*

I. INTRODUCTION

In telecommunications, a voice communication is a system of rules that allow two or more entities of a communications system to transmit voice via any kind of variation of a physical quantity. These are the rules or standard that defines the syntax, semantics and synchronization of voice communication and possible error recovery methods. Protocols may be implemented by hardware, software, or a combination of both.

Speech coding techniques use well-defined formats for coding and decoding voice dada.[1] Each sampled voice has an enough meaning intelligible for the receiver to elicit a response from a range of possible responses pre-determined for that particular situation. The specified behaviour is typically independent of how it is to be implemented. Speech communication protocols have to be agreed upon by the parties involved.[2] To reach agreement, a protocol may be developed into a coding standard. A coding standard describes the same for computations, so there is a close analogy between speech communication and speech coding standards: protocols are to communications what coding standards are to computations.[3]

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. Voice communications is an indispensable part of network communication 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 when used in VoIP.

This paper is organized as follows. Section II describes various existing speech coding techniques. Section III describes Data Sources and Methodology. Section IV describes The Speech Coders And Its Features. Section V describes The Comparative Study. Section VI describes the proposed encoding and decoding algorithms. Finally, Section VII concludes the paper.

II. EXISTING SPEECH CODING TECHNIQUES

Statistics speech coding has undertaken the mean to provide sophisticated voice compression on the strategic aspects, sampling activities and synthesis tactics used by modern speech coders.[4] This study also collected information on the involvement of semantic speech coding in usual voice coding chains. The survey's questions address the following themes: speech analysis strategies and monitoring, synthesis structure, operational activities, relocation voice sample activities, coding activities, coding practices and relationships with real time propagation, advanced technology use, coding/synthesis/transmission/technological innovation, speech production performance management, speech coding management, speech product and voice structure, technological support programs, and obstacles to innovation.[1]

To increase the analytical potential speech coding, the speech coders always had plans to combine the data obtained from the original speech with data from Statistical analysis of the speech or administrative data through the algorithms like GMM and HMM. Usual speech coding may combine the information collected through its basic sampling and coding with information collected from statistically collected algorithmic sources of the input data, including its semantics. The information compiled from this analysis will be used by the speech coding to better understand the impact of strategy and innovation decisions and operational adaptations on the sophisticated and robust well

formed final speech code, as well as to develop standards to help VoIP applications to improve their productivity and competitiveness.[5]

III. DATA SOURCES AND METHODOLOGY

The target population for this study of Innovation and coding Strategy was defined to meet information needs at different levels of speech coding standards like Wide band speech coding and narrow band speech coding. These standards are currently industry detail for the core survey content (coding strategies), for a module on VoIP application chains and for a module on innovation. The coding strategy for this study was limited to standards within the following 14 sectors defined according to the ITU-T Perceptual evaluation of speech quality (PESQ) tool Sources (P.862 (02/01):

Narrow-band speech coding

- G.723.1, G.726, G.728, G.729, iLBC and others for VoIP or videoconferencing
- Full Rate, Half Rate, EFR, AMR for GSM networks
- SMV for CDMA networks

Wide-band speech coding

- G.722, G.722.1, Speex and others for VoIP and videoconferencing
- AMR-WB for WCDMA networks
- VMR-WB for CDMA2000 networks.

IV. THE SPEECH CODERS AND ITS FEATURES

It is well understood that a substantial proportion of sampled signals the prediction gain also depends on the sampling frequency , that is meant to be distributed to eligible sample frames. The extent of this “prediction” of speech signal grain has been a matter of speculation for some time. Two recent surveys shed further light on the matter.

The diversion ratio (proportion of speech prediction gain “prediction” to the synthesized speech) has been estimated by several researchers in the past by matching Sample Survey data of their codes on input speech samples with the existing speech coding methods “offtake.” The former tell us how much gain speech coding are gaining for speech compression.. The difference is a rough estimate of the extent of diversion.

Based on this method, the estimated diversion ratio in performance of hybrid coding was around 54 per cent in 2010-15, compared to other coding standards lake wave form coders and parametric coders. Needless to say, this is an alarming figure. To deliver good quality efficient speech coding schemes, hybrid coders has to compromise between waveform and parametric coders. Hybrid coding has the lowest diversion rate (around 7 per cent); the rate was well below the statistical average in the other coding also (around 25 per cent in each case). By contrast, the estimated diversion rates ranged between 85 and 95 per cent in wave form and parametric coding standards. These estimate, suggest a comprehensive breakdown positive to the hybrid coding towards real time VoIP.

The G.722 uses a lossy sub-band ADPCM algorithm with a sampling rate of 16kHz . It works on bit rate of 64 kbit/s (comprises 48, 56 or 64 kbit/s audio and 16, 8 or 0 kbit/s auxiliary data) and will have 14 bits/sample with a latency of 4ms.it support constant bit rate (CBR) and does not support variable bit rate(VBR). The G.722.1 uses a Modulated Lapped Transform, (based on Siren Codec), Lossy algorithm. It works on bit rate of 24,32 kbits/sec and will have 16 bits/sample with a latency of 40ms.it support constant bit rate (CBR) and does not support variable bit rate(VBR). The G.722.1C uses a Modulated Lapped Transform, (based on Siren Codec), Lossy algorithm. It works on bit rate of 24,32 kbits/sec and will have 16 bits/sample with a latency of 40ms.it support constant bit rate (CBR) and does not support variable bit rate(VBR).

The G.722.2 (AMR-WB) uses a multi-rate wideband ACELP, Lossy algorithm with a sample rate of 16KHz. It works on bit rate of 6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85 kbit/s and will have 14 bits/sample with a latency of 25ms.it support constant bit rate (CBR) and variable bit rate(VBR). The G.723 uses a ADPCM, Lossy algorithm with a sample rate of 8KHz. It works on bit rate of 24, 40 kbit/s and will have 13 bits/sample..it support constant bit rate (CBR) and does not support variable bit rate(VBR). The G.723.1 uses a MP-MLQ, ACELP, Lossy algorithm with a sample rate of 8KHz. It works on bit rate of 5.3, 6.3 kbit/s and will have 13 bits/sample with a latency of 37.5 ms .it support constant bit rate (CBR) and does not support variable bit rate(VBR).

The G.726 uses a ADPCM, Lossy algorithm with a sample rate of 8KHz. It works on bit rate of 16, 24, 32, 40 kbit/s and will have 13 bits/sample with a latency of 125 µs..it support constant bit rate (CBR) and does not support variable bit rate(VBR). The G.728 uses a ADPCM, Lossy algorithm with a sample rate of 8KHz. It works on bit rate of 16 kbit/s and will have 13 bits/sample with a latency of 0.625 ms. It support constant bit rate (CBR) and does not support variable bit rate(VBR). The G.729 uses a CS-ACELP, Lossy algorithm with a sample rate of 8KHz. It works on bit rate of 8 kbit/s and will have 11.8 bits/sample with a latency of 15ms.It support constant bit rate (CBR) and does not support variable bit rate(VBR).

The iLBC uses a block independent linear predictive coding Lossy algorithm with a sample rate of 8KHz. It works on bit rate of 15.2 kbit/s for 20 ms frames, 13.33 kbit/s for 30 ms frames. It

support constant bit rate (CBR) and does not support variable bit rate (VBR).

If this tentative line of explanation is correct, two conclusions can be drawn. First the speech coding standards add to growing evidence of steady improvements in the voice coding robustness in recent years. There is still a long way to go in achieving anything like acceptable levels of functionality, especially under the minimum bit quota, but recent progress shows that the semantic speech coding tried in MPEG-4 is not a "lost cause" — far from it. Second, one thing that really helps to prevent high processing requirement is to give coding a strong stake in the system (large quantities, low prices), and make sure that they are clear about their entitlements. That has already happened, to a large extent, with the recent coding techniques envisaged the one like in iLBC: it has become much simpler to make real time coding, because they know its coding due and clamor for it if need be.

The recent turnaround of the speech coding like G.729(used in videoconferencing also built largely on this simple insight, as well as on the related fact that broad coverage strengthens coding standards for a sophisticated VoIP. There is an important lesson here for the proposed by the recent developments of speech coding standards.

V. THE COMPARATIVE STUDY

The study starts with comparative study of two base algorithms : G.723.1 and iLBC

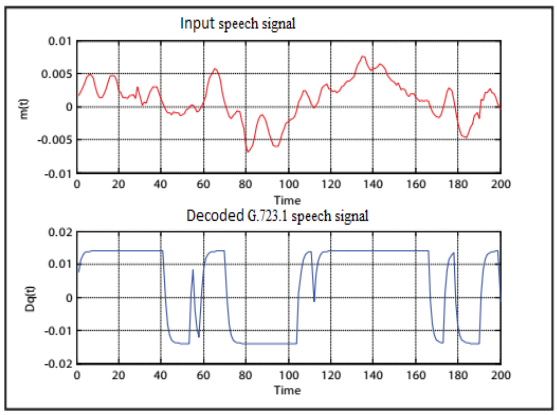

Fig.1 Encoding and Decoding for G.723.1

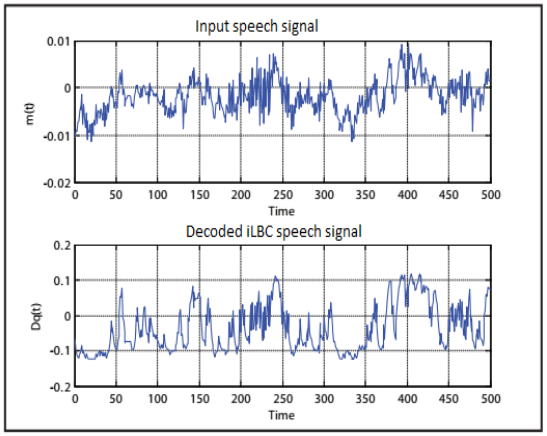

Fig.2 Encoding and Decoding for iLBC

MATLAB simulation of the input voice for G.723.1 and iLBC coders have been plotted graphically (Fig.1-Fig.2). iLBC reproduces the signal more closely to the original signal as compared to other coders We note that as the bit-rate goes down, the computation requirements increases highly for different bits used. This is the motivation for the proposal of semantic based speech code which is been explain in the next section. LP estimation for iLBC is depicted in fig-3.This introduces a delay as well as an increase in the cost of implementation. However, for equal number of bits used bandwidth in G.723.1and iLBC is reduced highly than waveform coders, making them most suitable in bandwidth scarcity situations

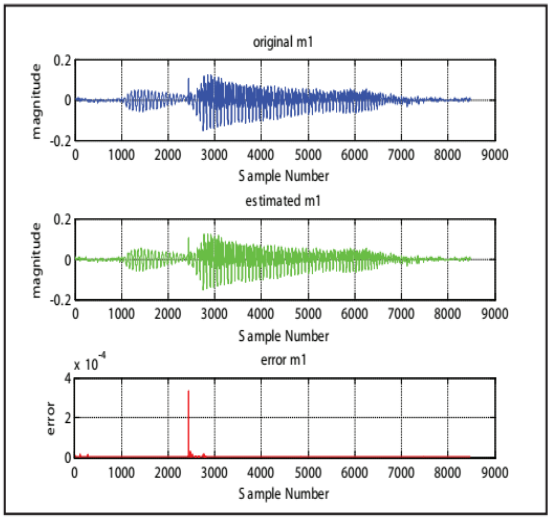

Fig 3. LP estimation for iLBC

6. THE PROPOSED METHOD

The proposed algorithm has 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.

Step 1: with LPC $e(n) = x(n) - \sum_{k=1}^{p} \alpha_k \, x(n-k)$,. The smaller the error , we have better set of predictors α_1 , α_2 , α_3 ,---------- α_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 $\alpha_k' s$,. Since this is not computational efficient , 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: 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 centroid 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 centroids to match the spoken word with the present database[3].

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 follows.

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 synthesized with the help of GMM

VI. CONCLUSIONS

In this paper, a novel semantic based speech compression methods which achieve the best possible speech quality low bit rate, with constraints on complexity and delay. In this paper, three two categories of speech coders were studied using MATLAB: Narrow-band speech coding and Wide-band speech coding.

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.

Speech coding algorithms are improving day by day to address the issues of speech communication standards. Even though this issue is addressed and solved, the VoIP industry demands more low bit energy efficient speech codes.

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] Dhinesh Babu L.D, P. Venkata Krishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments", Applied Soft Computing 13 (2013), pp.2292–2303.

[5] Matthias Schmidt,Niels Fallenbeck,Matthew Smith,Bernd Freisleben,"Efficient Distribution of Virtual Machines for Cloud Computing",Proceedings of the 2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing,IEEE Computer Society Washington, DC,(2010), pp.567-574

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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 .

[13] 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.

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

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

[16] 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

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

[18] Unknown,http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[19] 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

[20] Unknown,http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

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

[22] 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.