CONTEXT-BASED ERROR RECOVERY TECHNIQUE FOR GSM AMR SPEECH CODEC

Vinay MK and Suresh Babu PV
vinaymk@emuzed.com, pvsuresh@emuzed.com
Emuzed India Pvt Ltd.
Bangalore – 560 008, India

ABSTRACT

GSM AMR speech codec being used for both Internet and mobile networks, robustness to both frame erasures and random bit errors assumes significance. This paper proposes a new context-based error recovery technique for the CELP-based speech codec accomplishing recovery of erased frames, updating decoder state during erasure spells and reliable estimation of codec parameters in case of bit errors. Previous error concealment techniques do not adequately make use of the context in which concealment is being done. The proposed error concealment technique is intended to retrieve and use contextual information for better performance. The method is solely receiver based, uses no look ahead, makes use of implicitly available codec parameters and buffers for parameter estimation and is hence computationally efficient. Segmental Itakura-Saito measure and MOS scores are used to compare the output speech quality of the proposed technique with those of the basic techniques as recommended by the standard.

1. INTRODUCTION

Voice is one of the major communication applications connecting Internet and mobile networks. With each of these networks being characterized by frame erasures and random bit errors respectively, an integrated approach to receiver based error recovery becomes important. There have been many error recovery techniques suggested for ADPCM and other waveform codec based systems [2]. But these methods are not suitable for model-based codec systems, as they do not update the decoder states during errors. Codec based error recovery technique of Cupermann et al [3], attempts to answer this problem. With this method, the contextual information is in the form of speech classification into distinct classes, transmitted along with encoder data affecting interpretability. With CELP-based codecs error recovery mechanism becomes a necessity, to avoid unstable model state during and after error spells.

Speech characterized by its quasi-stationarity can be split-up into segments called phoneme classes or contexts. In a CELP-based decoder, model parameters are estimated according to the recovery technique used during/after the error spells. In this paper an attempt has been made to reliably estimate the model parameters during and after the error spell based on neighboring contextual information. The proposed error recovery technique can be used for both complete frame erasures and random bit errors. Interoperability issues confines it to be receiver-based solution, while low-delay requirements rules out look ahead, to avoid the delay and the need to operate in real-time on mobile/handheld devices necessitates it to be a computationally efficient method. The speech quality after error recovery is quantified using Itakura-Saito (IS) measure [4] on a frame-to-frame basis and MOS scores for entire speech files.

2. SYSTEM OVERVIEW

Global System for Mobile communication's Adaptive multirate (GSM AMR) speech coding standard has emerged to be a dominant technology for voice applications in both 3G and IP networks [5]. GSM AMR is a CELP based speech codec on 8KHz speech signal, capable of operating at eight different bit rates (4.75Kbps – 12.2Kbps). Voice activity detection is employed to facilitate discontinuous transmission, coupled with a comfort noise generation block at the decoder. The short-term correlation in the speech as captured by Linear Prediction (LP) analysis is transmitted in the form of quantized residuals of Line Spectral Frequencies (LSF). The excitation signal is constructed using pitch dependent adaptive codebook vector, residue dependent algebraic codebook vector and their gains. The adaptive codebook vector is constructed by repeating the delayed and filtered version of the previous excitation vector.

The proposed error recovery technique is applied to a real-time voice application on IP, employing GSM AMR speech coding standard. In voice applications over IP, more than one frame is packed together to reduce transmission overhead. Each packet length can be from 40 to 80 ms, encapsulating 2 to 4 speech frames. In the IP

scenario a delayed packet is as good as lost, hence for IP specific applications packet loss/delay implies an erasure spell of 40-80ms. For simulations encoder bit stream of every two frames is grouped into a packet, with no interleaving. The proposed error recovery technique assumes no look ahead, and hence is applicable to non-IP systems as well. Incase of packet loss or bit error the decoder will estimate the codec parameters and suitably updates the decoder buffers for better perceptual quality of the speech.

3. BASIC ERROR RECOVERY

In CELP-based speech codecs speech quality degrades drastically with errors in model-parameter estimation. The missing packets and associated errors in parameter estimation makes error recovery technique a necessity for packetised speech networks. The GSM standard recommends estimating codec parameters for erroneous frames based on previous history. LSF vector V_n for the erroneous frame is given by:

$$V_n[j] = (1-\alpha)\,M[j] + \alpha V_{n-1}[j], \quad j = 0..9 \qquad (1)$$

Where M is the constant mean LSF vector and V_{n-1} is the LSF vector of the last sub-frame and $\alpha < 1$ represents the adaptation constant. The adaptive codebook integer delay is given by the integer delay of the last sub-frame while fractional delay is set to zero. The adaptive codebook (pitch) gain, $g_p^{(n)}$ for the erroneous frame is given by:

$$g_p^{(n)} = g_p^{(n-1)} P[State] \qquad (2)$$

Where $g_p^{(n-1)}$ is the pitch gain of the last sub-frame and P [State] represents the decay rate. The index *State* represents the current position in the error recovery state machine, and has a maximum value of six. The algebraic codebook gain, $g_c^{(n)}$ for the erroneous frame is given by:

$$g_c^{(n)} = g_c^{(n-1)} C[State] \qquad (3)$$

Where $g_c^{(n-1)}$ is the code gain of the last sub-frame and C [State] represents the decay rate. The decay rate is independent of the type of speech segment currently being concealed. The drastic reduction in amplitude, during voiced segments, makes it annoying even for short erasure spells.

4. CONTEXT BASED ERROR RECOVERY

The codec parameters characterize the speech segment they represent and hence are highly context dependent. Speech segments has been traditionally classified into voiced, unvoiced, transition and silence contexts. More specific context determination comes at the cost of increased complexity. Now making use of the short-term stationarity of speech signal we propose to integrate contextual information in estimating the codec parameters, as an error concealment technique. Unlike other methods, a soft decision is made based on the voicing level for every speech frame, instead of classifying it into voiced/unvoiced classes. We define a parameter *VL,* which indicates the voicing level of the correctly received speech frame on a scale of 0-8. A value of 8 indicates strongly voiced, while 0 indicates completely unvoiced region. The *VL* is based on all the four sub frames of a frame, and no additional computation needs to be done as it is implicitly computed in the standard code. *VL* value of the last correctly received frame alone is used for all error recovery techniques.

4.1. LSF vector estimation

Speech phonemes and voicing levels will have a direct correlation, there will be a high degree of correlation between the mean LSF vectors (which represent a speech segment) and their voicing levels. Simulations confirmed the high degree of correlation between the mean LSF vectors and their voicing levels. The mean LSF vectors classified according to three values of *VL* from a speech file of about 10sec are plotted in Fig 1. Simulations shown that there is substantial variation in mean LSF for different *VL,* while the constant standard mean vector strikes a compromise by following *VL = 4* level.

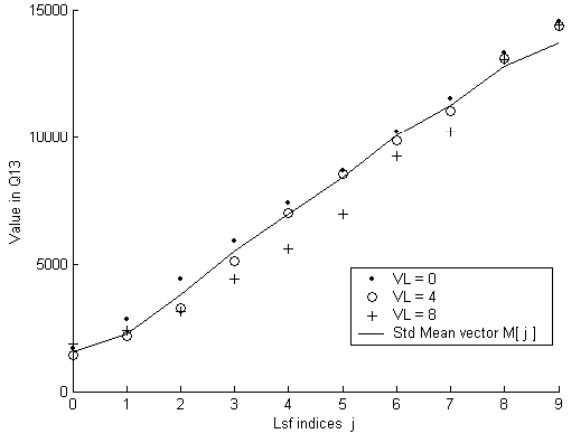

Figure 1. Distribution of the Mean LSF vectors as a function of voicing levels.

Hence defined a Mean LSF vector set, adaptively updated for every *VL*. The adaptive nature of modification of mean LSF vectors will bring in features of speaker also into consideration for better perception quality. For every correct frame received the mean LSF vector set is updated as:

$$M[\mathrm{VL}][j] = \beta M[\mathrm{VL}][j] + (1-\beta)\, V_n[j] \quad for\ j = 0..9 \tag{4}$$

Where β is the adaptation constant and *VL* is its voicing level. In the proposed technique the LSF vector, V_n, for the lost frame is given by:

$$V_n[j] = (1-\alpha) M[VL][j] + \alpha V_{n-1}[j] \quad for\ j = 0..9 \tag{5}$$

Where *VL* is the voicing level of the last correctly received frame.

4.2. Gain Fading

Reconstructed frames during erasure spells are subjected to monotonous decay, to avoid spurious noise due to concealment efforts during long erasures. The pitch and the code gains are subjected to gradual fading for successive lost sub-frames. From the experiments it has been observed that the uniform decay rate for all contexts may not be really suitable. Excessive decay during high voiced regions (middle portion of a vowel) leads to hollowness in the reconstructed speech.

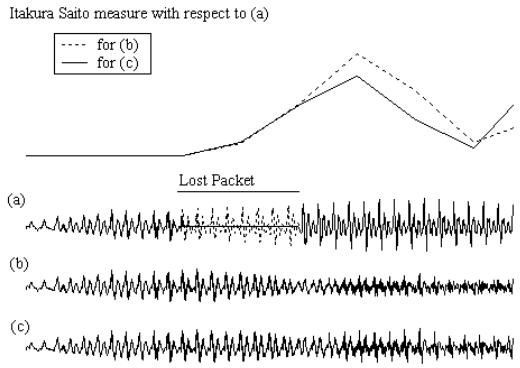

Figure 2. Effect of different gain fading schemes measured using the Itakura-Saito measure. a) Original decoder output with no error, lost packet is superimposed in dotted b) decoder output with standard error recovery technique, using uniform fading rate c) decoder output with proposed error recovery technique, which uses lower decaying rate constant for voiced regions.

A Typical voiced segment is as long as 140ms. Even with a small loss of 40ms speech, one can notice substantial decay of amplitude in the reconstructed speech with uniform fading, in Figure 2. Hence we suggest a gain fading scheme, which depends on the voicing level (of the last correctly received frame). In the proposed technique the adaptive codebook (pitch) gain, $g_p^{(n)}$ is given by:

$$g_p^{(n)} = g_p^{(n-1)} P[VL][State] \tag{6}$$

Where $g_p^{(n-1)}$ is the pitch gain of the last sub-frame and P [*VL*][State] represents the decay rate, which depends upon the voicing level. While the algebraic codebook gain, $g_c^{(n)}$ estimation remains as in (3).

4.3. Minimizing Error Propagation

Frame loss or bit errors lead two types of errors (deviation from ideal output), firstly the error in estimating the codec parameters during the erasure spell, secondly improper updating of codec states leading to propagation of error beyond the erasure spell. The propagation of error is significant in the case of codecs with backward prediction of parameters. Hence the error recovery technique is required to estimate the codec parameters reliably during error spells and also to suitably update the decoder states (history buffers) to minimize error propagation.

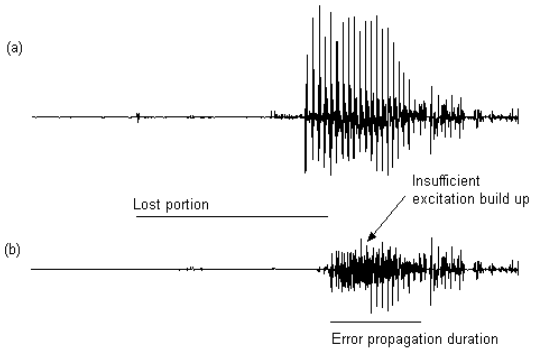

Figure 3. Error propagation in excitation buffer due to its improper updation

In the GSM AMR decoder the LSF vector and the algebraic codebook gain are derived by first order prediction of the residuals. The adaptive codebook excitation vector is obtained by filtering the suitably delayed version of the previous excitation buffer. The error propagation in the case of codebook gain is limited to first sub-frame; LSF vector error is limited to first

frame. But the error due to excitation buffer update can extend to several frames depending upon the context.

Simulations shows that this effect is severe when the starting portion of a vowel segment is lost and the error propagation length beyond erasure spell, is directly proportional to the voicing level of the lost frame(s). Figure 3 further substantiates these conclusions. Incase of voiced frames adaptive codebook excitation has significantly high amplitude. When the starting portion of a voiced segment is lost the build up in the excitation amplitude is affected. Hence depending upon the *VL* of the correctly received frame and the current excitation energy, its excitation buffer needs to be scaled up. Substantial improvement in the perceptual quality and reduction in perceived hollowness during and after erasures by using the proper scaling of excitation buffer as explained above and the results shown in Figure 4.

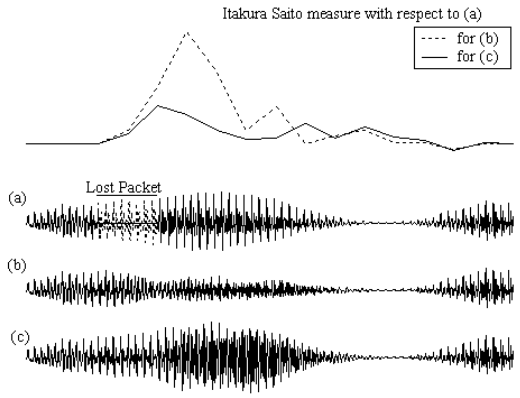

Figure 4. Excitation scaling to minimize error propagation, as measured using Itakuara-Saito measure. a) Original decoder output with no error b) decoder output with standard error recovery technique using no post error manipulation of decoder buffers c) decoder output with proposed error recovery technique using excitation scaling in case of voiced segments VL > 4.

5. SIMULATION AND RESULTS

In the context-based error recovery technique, three basic methods are proposed, by which the contextual information can be incorporated. Ten phoneme rich speech files from the TIMIT database [6] of the speaker "mbgt0" are concatenated and down sampled to 8KHz to produce a test speech file of about 32 seconds (TSF-1). T21 test vector of about 8 seconds from ETSI [7] is chosen as the second test speech file (TSF-2). These test speech files are coded using GSM AMR encoder at 7.95 Kbps with DTX enabled. Error sequence made up of 0's and 1's are generated from a uniform distribution with approximately 5%, 10% and 20% of them being 0's, representing frame erasure rate (FER). A packet loss results in loss of two consecutive frames, each of 20ms duration. Incase of bad speech frames all the parameters are estimated using the error recovery technique. Table 1 shows the mean opinion scores (MOS) as obtained for the test files after applying the standard (REF) and the proposed context-based error recovery (CER) techniques independently. Results shows that the proposed error recovery technique performs better than standard technique even at very high error rates.

FER		REF	CER
5%	TSF-I	3.27	3.39
	TSF-II	3.68	3.55
10%	TSF-I	2.08	2.25
	TSF-II	2.27	3.43
20%	TSF-I	1.64	1.76
	TSF-II	1.68	1.80

Table I. MOS scores for various FER with standard (REF) and proposed error recovery (CER) techniques as applied to the two test files.

6. CONCLUSIONS

The perceptual quality of the proposed context based error concealment algorithm shown a good improvement compared to the standard algorithm for a variety of speech data and various frame erasure rates ranging from 5 to 20%. The speech quality measures MOS and Itakura-Saito measure shown superior performance when applied to two different speech files at different Frame Erasure Rates (FER). Except in one case with TSF-II and FER = 5%, MOS scores show degradation. The proposed technique can be extended to other CELP based speech coders.

7. REFERENCES

[1] 3G TS 26.090: "AMR Speech Codec; Transcoding Functions version 4.0.0"

[2] C. Perkins, O. Hodson, V. Hardman, "A Survey of Packet-Loss Recovery Techniques for Streaming Audio." IEEE Network Magazine Sept/Oct 1998.

[3] A Hassan and V. Cuperman, " Reconstruction of Missing Packets For CELP Based Speech Coders". Proc ICASSP 95.

[4] D.G.Jamieson, L.Deng, M.Price, Vijay Parsa and J. Till, "Interaction of Speech Disorders with Speech Coders: Effects on Speech Intelligibility", Proceedings of ICSLP 96.

[5] Wireless Multimedia Forum, Recommended Technical Framework document for streaming media over wireless networks, version 1.0.

[6] TIMIT Speech Database," http://www.mpi.nl/world/tg/corpora/timit/timit.html".

[7] 3G TS 26.074: " AMR Speech Codec Test Sequences version 4.0.0"