14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A VOICING DECISION DRIVEN FORWARD ERROR CORRECTION SCHEME FOR VOICE OVER IP APPLICATIONS

Dawn A Black and Mark B Sandler

Centre for Digital Music
Queen Mary University of London
email: dawn.black@elec.qmul.ac.uk

ABSTRACT

This paper examines the performance of a new packet loss concealment (PLC) schemes under bursty loss conditions. We confirm that burst losses cause higher distortion than equal rates of single packet loss, and introduce a new Forward Error Correction (FEC) and PLC combination that reduces artifacts typical of PLC schemes. FEC informs the receiver of the pitch and voicing of lost speech, combined as a single metric. A new PLC scheme combines ideas of repetition and interpolation with new approaches made possible by FEC to create loss concealing speech. The performance of the new scheme is assessed through comparison with the existing standard, and the increase in bit rate incurred by FEC is quantified.

1. INTRODUCTION

The use of internet telephony is now fairly widespread, but quality of experience (QoE) issues still remain. Internet telephony works fine over a LAN, but it is far from being able to provide a high QoE for those with dial-up modems at home. Data networks are also unable to resolve fault conditions quickly enough for voice traffic to be unaffected. Open Shortest Path First [?] converges in up to three minutes, and the Spanning Tree Protocol [1] in thirty seconds, but even that is too long for voice signals. Additional network characteristics such as jitter, delay and buffer overflow combine with these problems to form packet loss. Methods which attempt to conceal such loss are known as Packet Loss Concealment (PLC) methods.

PLC methods can be divided into two groups, receiver based and transmitter based. Receiver based methods rely on exploiting the short term stationarity of speech in order to estimate the missing signal. Techniques may utilize the information in previously correctly received 'good' packets only (*a priori* knowledge based PLC). Or, since the loss of a packet is often detected by receiving the next good one, the information in both past and future correctly received packets can be used (*a posteriori* knowledge based PLC). From here on, audio frames contained in successfully received packets will be referred to as good-frames, and we allow one 20 ms frame per packet.

A priori receiver based PLC techniques include silence insertion, where silence is substituted for the missing speech; frame repetition, where the missing data is concealed by repetition of the last correctly received packet; and other waveform substitution methods such as the G.711 Reverse Order Replicated Pitch Period [2]. A priori methods are generally based on frame repetition.

A posteriori receiver based PLC techniques tend to be interpolative in nature and generally outperform *a priori* PLC methods, justifying the additional delay [3]. Applications can be frequency domain [4], or time domain [5].

Transmitter based recovery schemes attempt to code the speech with loss in mind. Such schemes often work by introducing redundancy at packet level and these are collectively known as FEC schemes.

The overall performance of PLC schemes is highly dependant on the prevailing network characteristics which tend to be bursty, meaning that consecutive packets are lost. Bolot *et al.* showed in [6] network loss statistics with an average of 20% loss in bursts of one or two packets. However burst losses of up to 15 packets were experienced and the probability distribution of the burst length was shown to decrease geometrically away from a burst loss of one. Loguinov and Rahda [7] show similar network characteristics, with on average 90% of loss occurring in bursts of one or two packets, but burst lengths of 20 or more packets were experienced. It is therefore quite probable that burst losses of more than two packets will occur.

Under high loss conditions *a priori* PLC schemes based on frame repetition suffer from two notable problems: 1) Unnatural harmonic artifacts occur as the pitch in successive concealing frames is constant. 2) There is no guarantee that the lost frame had similar characteristics to the repeated frame. Under similar high loss conditions, *a posteriori* PLC methods which utilise interpolation may produce audible artifacts due to excessive smoothing of formants and energy. Both methods are unable to conceal the loss of a transition frame, in which the speech makes a transition from one voicing type to another, or between phonemes. Some effort has been made to suppress these artifacts by attenuating consecutive concealing frames, but this approach inevitably results in silence insertion for extended periods of loss. For a PLC method to be able to cope with bursty packet loss while avoiding the problems associated with repetition and interpolation, some element of FEC, although it increases the bit-rate, is unavoidable.

In this paper we introduce a novel FEC scheme partnered with a new *a posteriori* PLC method. We focus on minimising errors due to repetition and interpolation. This paper is organised as follows. Section 1.1 describes the ITU-T G.711 standard for PLC used for comparison purposes, section 2 describes the new FEC scheme and section 3 the partner PLC scheme. Section 4 details the testing methodology and results are presented in section 5.

1.1 G.711 PLC method

G.711 [2] is the ITU recommended loss concealment method. It is an *a priori* pitch repetition based PLC scheme. When loss occurs, concealment is effected through repetition of one or more pitch periods from the previous frame(s). The pitch and voicing of the lost frame is assumed to be the same as that for the previous 48.75 ms of speech. Repetition based artifacts are alleviated, but not eradicated, by changing the pitch periods selected for repetition. The loss concealing speech is also attenuated at a rate of 20% per 10 ms loss, starting after 10 ms total loss, so that after 60 ms the signal is zero. A triangular windowed OverLap Add (OLA) is used to smooth between both the real signal and loss concealing speech, and consecutive loss concealing frames. The length of the OLA depends on both the pitch period and the length of the erasure. For short, 10 ms erasures, a quarter pitch period window is used. For longer erasures the window length is increased by 4 ms per 10 ms of erasure, up to a maximum of 10 ms. The performance of our PLC scheme is compared to that of G.711.

2. A NEW FEC SCHEME

2.1 Voicing type (ϕ)

Speech can be classified as having voiced ($\phi = v$) sections which possess a clear pitch, and unvoiced ($\phi = u$) sections which are more noisy in nature. In its simplest form, voicing driven PLC replaces missing voiced frames with a signal with periodic characteristics, and missing unvoiced frames with random noise. Reliable estimation of the voicing content of a lost frame is therefore fundamental.

A posteriori and *a priori* PLC schemes infer the voicing type of the lost frame from surrounding good-frames, and thus are not always capable of predicting the voicing of a lost frame correctly, resulting in loss concealing speech with incorrect voicing characteristics. Our FEC scheme ensures the receiver has reliable information about the voicing content of lost frames by embedding the voicing type of k previous frames into the packet containing frame m. The voicing information is not transmitted as a parameter in its own right, but is combined with the pitch information as described next.

2.2 Pitch period information τ

The pitch period τ (in samples per period) of speech is an important parameter and synthesised speech with incorrect or constant pitch is perceptually disturbing. Approximating the pitch of lost frame n as τ_n, can help conceal short errors [2] [4] but is not capable of concealing extended periods of loss. For this reason our FEC scheme ensures the receiver knows the pitch of the missing speech by repeating τ of the previous k frames in the packet containing frame m. Voiced frames will have a clear pitch period greater than 1, we assign $\tau = 0$ to silent frames, and $\tau = 1$ to unvoiced frames. Hence the voicing of a lost frame n can be inferred from τ_n. Having accurate pitch information about lost frames due to FEC means we avoid artifacts arising from excessive repetition. Also, in the event of the loss of a transition frame somewhere within a burst loss we are able to make the transition from one type to another at the correct frame.

2.3 Bit-rate

A pitch and voicing detection function was applied to the test set described in section 4 to give the probability density function and range of τ. The range of τ was found to be from $\tau_{min} = 0$ to $\tau_{max} = 375$ samples. Extending this to $\tau_{max} = 400$, Huffman coding [8] assigned a maximum of 12 bits to the least probable pitch values, and a minimum of five bits to the most probable. On average τ required 5.85 bits per frame. The overall increase in bit-rate due to FEC is dependant upon the number of previous frames of redundancy k inserted into the packet containing frame m. In other words, the expected loss k must be decided in the encoder, and compensation for this made by embedding τ for k past frames into packet m. This determines the overall increase in bit-rate. It has been shown that this is best decided through use of adaptive control mechanisms that react to the prevailing network conditions [6].

3. A NEW PLC SCHEME

3.1 Good-frame Selection

The FEC scheme described in the previous section is now applied to the PLC scheme shown in figure 1. The loss concealing speech $\hat{S_n}$ is based on *either* the last S_{n-1} or next S_{n+b} good-frame (where $b = 1$ is the burst length for this example). This selection is made by the 'good-frame selection' block and can be either a one or two stage decision. The first stage is compulsory and is voicing driven, the type of the lost frame ϕ_n being supplied by FEC. We compare ϕ_n to ϕ_{n-1} and ϕ_{n+b}. There are three possibilities

1. All frames are of the same type. $\phi_n = \phi_{n-1} = \phi_{n+b}$
   a. **Voiced:** $\phi_n = \phi_{n-1} = \phi_{n+b} = v$. Good-frame selection proceeds to stage two, pitch based selection.
   b. **Unvoiced:** $\phi_n = \phi_{n-1} = \phi_{n+b} = u$. Frame S_{n-1} is selected by default.
2. $\phi_n = \phi_{n-1}$ OR $\phi_n = \phi_{n+b}$. The type of one frame only matches that of the lost frame and that frame is selected.
3. $\phi_n \neq \phi_{n-1}$ AND $\phi_n \neq \phi_{n+b}$. The type of neither frame matches that of the lost frame so S_{n-1} is selected by default.

Stage two, 'pitch based selection', is only implemented in the event of condition (1a). The good-frame is the one that has the pitch closest to that of the lost frame τ_n. If $|\tau_{n-1} - \tau_n| = |\tau_{n+b} - \tau_n|$, frame S_{n-1} is selected.

3.2 Pitch Period Based PLC

Having selected the good-frame, it is now repeated and altered to form the concealing speech. If the good-frame is unvoiced, the last quarter is simply repeated to conceal the lost frame. If the good-frame is voiced the number of periods in the lost frame is found $\eta_n = M/\tau_n$ where M is the number of samples per frame. The concealing speech is then formed by copying the last η_n pitch periods from the good-frame and re-sampling so that the correct pitch τ_n is achieved. There may not be enough pitch periods in the good-frame to replace the lost speech. This being the case and the good-frame being S_{n-1}, the last pitch period of S_{n-1} is repeated as necessary and appended to the start of $\hat{S}_n$ to make up the shortfall prior to re-sampling. If the good-frame is S_{n+b} the first pitch period of S_{n+b} is repeated and appended to the end of $\hat{S}_n$. Each consecutive lost frame is treated in the same manner and concealing speech is never used to conceal further lost frames.

3.3 Root Mean Square (RMS) energy interpolation and Overlap Add

This is the final stage in our PLC scheme. The RMS energy E is calculated for the last pitch period in frame S_{n-1} and the first pitch period in S_{n+b}. An expected and actual value of RMS energy for each pitch period in the concealing speech is then found; the expected value through linear interpolation. Each concealing pitch period is then scaled individually to meet the expected RMS value. This provides us with a smooth transition between good-frames but can result in over smoothing for long burst losses. An OLA operation with an overlap of $\frac{\tau}{4}$ between consecutive loss concealing frames, and an overlap of $\frac{M}{4}$ (where M is the number of samples per frame), is performed to smooth between frames.

4. TESTING

4.1 An annotated speech test set

A 16 KHz speech test set comprising five male and five female recordings (giving around 1500 frames of 20 ms in total) was taken from a selection of BBC audio books. A pitch and voicing detection algorithm was applied to each 20 ms frame to provide a value of τ per frame. Because both PLC schemes are pitch based it is important to note frames where the pitch is detected incorrectly as this may lead to large artifacts in the loss concealing speech. Therefore the accuracy of the

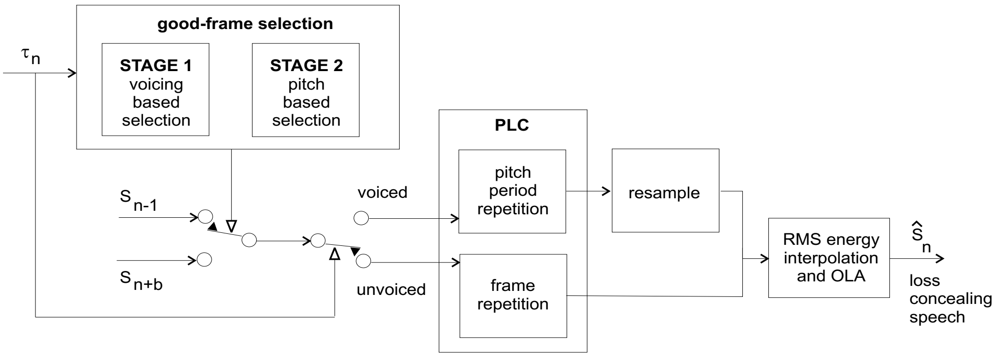

Figure 1: Block diagram depicting the stages of the new PLC method

pitch detection was noted 'by hand' to produce an annotated test set. Additional information such as the position of transition frames was also noted.

4.2 Distance measure

In [9] Yang *et al.* present a Modified Bark Spectral Distortion (MBSD) measure which is shown to exhibit a strong correlation to Mean Opinion Score (MOS) results. The MBSD quantifies the impact of distortion through the use of psychoacoustic models and compares the original and distorted signals on a frame-by-frame basis, producing a quality related (one being excellent, two good, three fair, four poor and five bad) MBSD value ρ for each frame. Thus a set of MBSD values $\Gamma = \{\rho_1, \rho_2, \ldots, \rho_{N^m}\}$ where N^m is the number of frames in speech database item m. MBSD values $\rho > 5$ are possible, indicating very bad distortion, and MBSD values are real values. An MBSD $\rho = 0$ indicates no distortion. Results are presented in terms of mean ρ, and the mean number of distorted frames (or frame errors) per second. The mean MSBD is calculated as

$$\frac{1}{s}\sum_{m=1}^{s}\frac{1}{N^m}\sum_{n=1}^{N^m}\rho_n \tag{1}$$

where $s = 10$ is the number of data items in the test set and N^m is the number of frames in test item m. The set of frame errors $\Omega \subset \Gamma$ where Ω is the set of all $\rho > 0$, allowing us to focus on the impact of loss (frames experiencing no loss will show no distortion). The mean number of distorted frames per second is given as

$$\frac{1}{s}\sum_{m=1}^{s}\left(\omega^m \Big/ \frac{N^m}{f}\right) \tag{2}$$

where f is the number of frames per second and ω^m is the number of items in Ω^m.

The original content of the lost frame is important as some speech is easier to conceal than others. In the event of loss, the annotated test set is used to provide probable cause of the distortion. Three classes are recorded:

1. **Transition error:** If loss occurs, and either the lost frame or the good-frame are transitional in nature.
2. **Wrong-type error:** If loss occurs and the good-frame chosen is of a different voicing type to the lost frame.
3. **Incorrect-pitch error:** If loss occurs and the pitch of either the lost or good-frame was incorrectly detected.

If there is a choice of outlier types priority is imposed according to the list order. The number of instances of each error type is recorded as Φ^m where $\Phi^m \subset \Gamma^m$ for test item m. We wish to show the contribution of each error type to the overall distortion and thus the mean of each set Φ^m and the size of each set ϕ^m (where ϕ^m is the number of ρ in each Φ^m) is given. The mean is defined as

$$\frac{1}{s}\sum_{m=1}^{s}\frac{1}{\phi^m}\left(\sum_{p=1}^{\phi^m}\Phi^m(p)\right) \tag{3}$$

4.3 Loss models

Loss models were created to simulate loss as a total of 10%, 20% and 30%. For each target total loss, frames were dropped in set bursts of $b = 1, \ldots, 5$ but at random locations, giving 15 loss models in total. Each speech item in the test set was subjected to loss according to each loss model. Speech frames not lost were not altered in any way.

5. RESULTS

The overall performance of G.711 and the new PLC algorithm in terms of MBSD is shown in figure 2. If all loss were concealed transparently, the mean MBSD would equal zero, and the mean frame errors per second would also be zero. Plot (a) shows the mean MBSD 2 for all frames. It is clear that the new PLC scheme reduces the overall MBSD significantly. The mean MBSD increases with loss rate and burst length, but the results for a low percentage total loss with long burst lengths are generally poorer than those for greater overall loss but shorter burst lengths (for example, 20% loss in bursts of 5 packets shows worse quality than 30% loss in bursts of 2 packets) . This confirms that long burst losses cause greater distortion than randomised loss. Plot (a) also shows that, as the overall loss and burst length increases, so does the difference between the performance of G.711 and the new PLC scheme. For example, for 30% loss with a burst length of one, the new PLC scheme shows a reduction in MBSD of about 0.01. Compared to 30% loss in bursts of five packets, we see that the new PLC scheme now shows an improvement of about 0.4. We can therefore conclude that the new PLC scheme is better able to conceal burst losses.

Plot (b) gives the mean frame errors per second 3. We can see that the number of frame errors (frames which experience loss) is approximately equal for all equal % loss tests. Since the mean MBSD is lower for the new PLC scheme, but the number of distorted frames is the same as for G.711, the conclusion is that the new PLC scheme produces lower distortion, and is closer to concealing loss transparently.

Plot (c) shows the mean MBSD for each error type. Both G.711 (black) and the new PLC (white) show similar levels of distortion due to incorrect-pitch outliers (black), as both are heavily dependant on accurate pitch information. Both also perform similarly in terms of lost transition frame (white) concealment. However, the new PLC scheme shows a clear reduction in terms of wrong-type errors (grey). The *a posteriori* approach and FEC allow an informed good-frame selection, reducing wrong-frame outliers. Improving the concealment of transition frame losses is the subject of further work.

Plot (d) shows the total number, over the entire speech data set, of frame errors occurring for each error type. It is very clear that the new PLC (right bars) scheme experiences almost no instances of wrong type errors (grey), a significant reduction from G.711 (left bars).

6. CONCLUSIONS

This paper presented a new combined FEC and *a posteriori* PLC scheme. The FEC piggy-backs redundant Huffman coded pitch and voicing information about previous frames onto current audio frames, thus ensuring the PLC scheme knows the pitch and voicing of a lost frame, requiring on average 5.85 bits per past frame. The PLC is repetition based and FEC enables the PLC to make an informed choice about the frame to repeat. The a posteriori nature of the PLC allows adjustment of the energy of the loss concealing speech through interpolation. When compared to the existing standard G.711 the performance of the new scheme was shown to perform better overall, and to be more robust to bursty loss conditions.

REFERENCES

[1] IEEE Standards Association, “Ieee standard 802.1d,” .

[2] “Packet loss concelament algorithm for use with ITU-T recommendation G.711,” ANSI Recommendation TL521-2000 AnnexB.

[3] J. Wang and J.D. Gibson, “Performance comparison of intraframe and interframe LSF quantization in packet networks,” in *Proc IEEE Workshop on Speech Coding*, 2000, pp. 126 –128.

[4] F. Bouteille, P. Scalart, and B. Kovesi, “Packet loss concelament using audio morphing,” in *Speech Processing Transmission and Quality Aspects Workshop*, 2003.

[5] J Tang, “Evaluation of double sided periodic substitution (dsps) method for recovering missing speech in packet voice communications,” in *Computers and Communications, IEEE Conference Proceedings*, 1991, pp. 454 – 458.

[6] J.C. Bolot and A. Vega-Garcia, “Control mechanisms for packet audio in the internet,” in *Proc. INFOCOM*, 1996, vol. 1, pp. 232 – 239.

[7] D. Loguinov and H. Radha, “Measurement study of low-bitrate internet video streaming,” in *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, 2001, pp. 281 – 293.

[8] D.M. Huffman, “A method for the construction of minimum redundancy codes,” in *Proceedings of IRE*, 1952, vol. 40, pp. 1098–1101.

[9] W. Yang, M. Benbouchta, and R. Yantorno, “Performance of the modified bark spectral distortion as an objective speech quality measure,” in *Proc. ICASSP*, 1998, vol. 1, pp. 541 – 544.

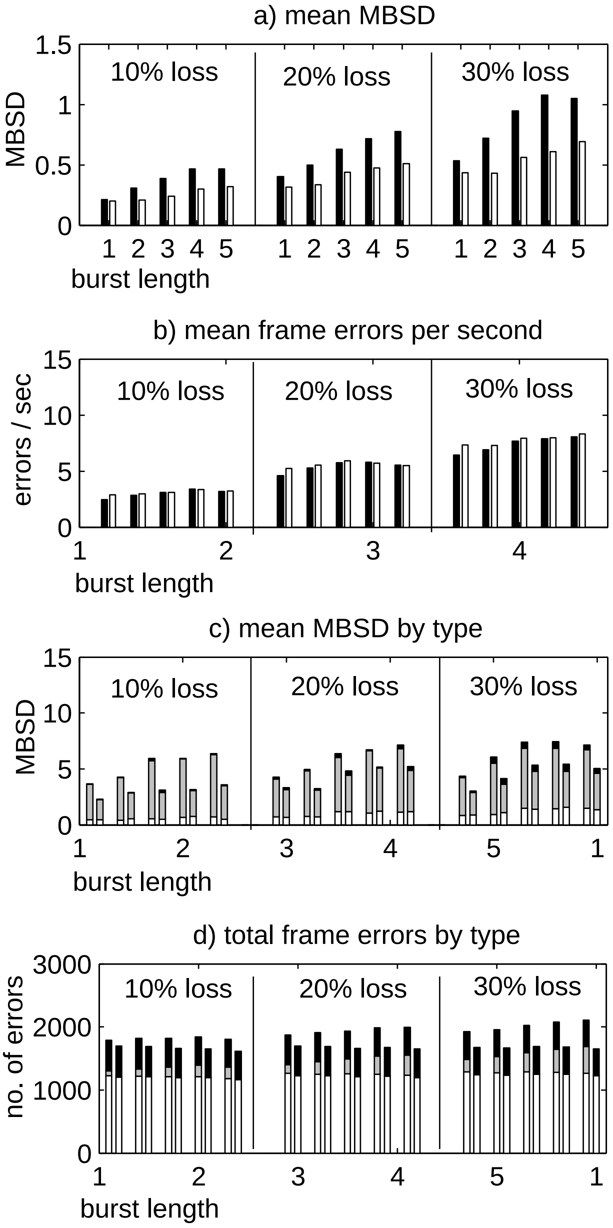

Figure 2: (a) Mean overall MBSD for G.711 (black) and the new PLC scheme (white). (b) Mean error frames per second for G.711 (black) and the new PLC scheme (white). (c) Breakdown of MBSD into (white) transitional errors, (grey) wrong type errors, and (black) incorrect pitch errors for G.711 (left of each pair) and the new PLC scheme (right of each pair). (d) Breakdown of total frame errors into (white) transitional errors, (grey) wrong type errors, and (black) incorrect pitch errors for G.711 (left of each pair) and the new PLC scheme (right of each pair).