PACKET LOSS CONCEALMENT USING AUDIO MORPHING

Franck Bouteille[1], Pascal Scalart[2], Balazs Kovesi[2]
[1] PRESCOM SA, Rue de Broglie, 22300 LANNION – France – Tel. : 33 2 96 05 38 27
[2] France Telecom R&D, 2, Av Pierre Marzin, 22300 LANNION – France
{franck.bouteille,pascal.scalart,balazs.kovesi}@rd.francetelecom.com

ABSTRACT

In packet switched networks (case of IP networks) no minimal threshold on rate is offered for the transport of packets. The available bandwidth depends on conditions of the traffic on the network (best effort strategy). In VoIP (Voice Over Internet Protocol) audio packets must be continuously played. When packets are lost or have arrived too late it gives place to chopped sound and to a degradation of the quality. In this paper the lost case is considered when a packet loss is recognized by the reception of the next packet. Generally concealment techniques do not use this received signal. We propose here an algorithm, which uses this information to improve the missing signal synthesis and allows maintaining an acceptable quality of the speech signal.
A comparison between this method and techniques recommended by the ITU-T for coders G.711 [3] and G.723.1 [1] will also be described.

1. INTRODUCTION

Packet communication techniques were originally designed for digital data transmission. In packet data networks, excess traffic leads to delays or loss in delivery of information. When a network is congested, packets are held in queues at entry points and switching nodes. In voice communication, on the other hand, long delays are intolerable and network delay budgets have strong influence on the design of packet voice systems. Beyond a certain limit, delayed packets are ignored by the receiving terminals or can be dropped at an intermediate node. Hence, congestion in packet voice systems leads to gaps in the received speech. The crackles in the reconstructed signal make the understanding difficult and the sound less natural. Even at low packet loss ratios such as 2 % these crackles are perceptible. At 10 % the intelligibility is always possible but requires a big concentration, and from 20-25 % understanding becomes impossible. The simplest way for handling gaps is to process them as silent intervals in the transmitted speech. But spurious silent intervals are audible and disturb listeners. To increase the tolerance of packet voice systems to lost packets some techniques have been developed. These methods are based on 2 types: Firstly "Waveform substitution" where the missing frames are replaced by another frame already received, using Pitch replication [6][3] or pattern matching [2], secondly "Model extrapolation" like the technique used in [1] and or as proposed in [5] where the model of previous received signal (eventually slightly modified) is used to generate the missing signal. These techniques do not use the *a posteriori* information of the next packet that indicates and detects the lost of one or several frames, which is not the case for the techniques proposed in [5] and [2]. However those last techniques are not adapted for long lost periods (>15ms) because of the non long-term stationnarity of speech signal. This *a posteriori* information is generally available because of the playout buffer management and real time network protocol [7].
The technique proposed in this paper uses the knowledge of the frame received after the last lost one, the models of the last received frames, and a model interpolation to synthesized the missing signal.

The remainder of this paper is organized as follows. Section 2 describes the principle of the proposed method and the way to implement it. Section 3 shows the results of the proposed method. After that, we confirm the effectiveness of our technique by comparing with G.723.1 method [1], G.711 method [3], DSPS method [6], an improved version of the algorithm described in [5] and the previous frame copy (PFC) method. Finally Section 4 concludes this paper.

2. PROPOSED METHOD

2.1. Principle

The proposed method uses a pitch period model and morphing techniques to estimate the missing signal. The principle of Morphing techniques is to transform, in a continuous way, a vector into another one even if these two vectors are very different.
Firstly we estimate the pitch period ($P0$) of the frame A (frame before the lost frame) and the pitch period ($P1$) of the frame B (frame after the lost frame).

Previous Frame Frame A	Missing Signal	Next Frame Frame B

Figure 1: context of lost

The method used to estimate the Pitch period is the Kobayashi method [4]. The strategy used to deal with unvoiced frames is presented in figure 2. Pitch values are searched in

$$2.5 \text{ ms} \le P0, P1 \le 15 \text{ ms} \qquad (1)$$

In the case where no voiced signal is detected for Frame A and Frame B, the concealment method used is a simple copy of the previous frame or a comfort noise generated to conceal the missing signal.

Else $P0$ and $P1$ are used to estimate the number of necessary intermediate blocks (*NbBloc*) and the size of these blocks (*SizeBloc*).

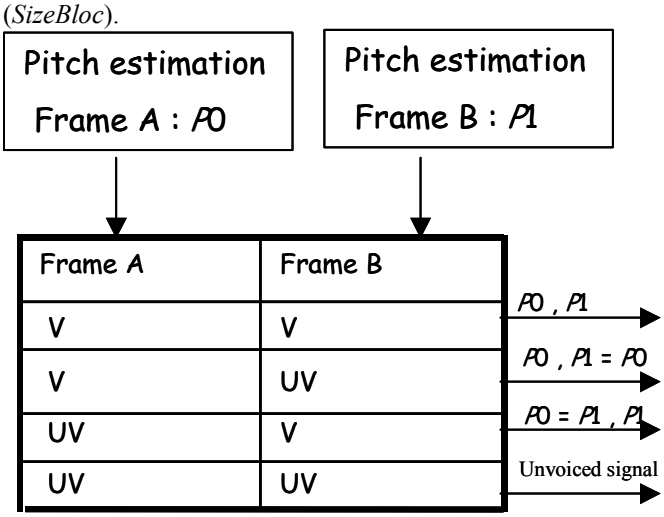

Figure 2: V (Voiced) / UV (Unvoiced) strategy

$$SizeBloc = \max(P0, P1) \qquad (2)$$

$$NbBloc = round\left[\frac{NbSampleLoss}{SizeBloc}\right] \qquad (3)$$

Once the number of blocks is estimated, we model the last pitch period vector (X0) of the Frame A (*ModP0*) and the first pitch period vector (X1) of the Frame B (*ModP1*). In order to have simple model parameters (*ModP0, ModP1*) we use a DCT (Discrete Cosinus Transform). DCT is estimated in 15ms vector (The maximum value of Pitch Detection, 120 samples at 8kHz of sample frequency), so that zero padding will be necessary in case of smaller pitch.

Intermediate blocks are used in order to transform, in a continuous way, the model vector *ModP0* to the model vector *ModP1* with linear interpolation of model parameters. In enhance versions of this morphing technique an improved interpolation method could be used to model the parameters evolution. In the current version of Audio Morphing, the intermediate blocks are computed in the following way:

$$Block_i(n) =$$

$$IDCT\left[ModP0(k) + i * \frac{ModP1(k) - ModP0(k)}{NbBloc}\right]_{120} \qquad (4)$$

$$0 \le i \le NbBloc - 1 \qquad 0 \le k \le 120 - 1$$

$$0 \le n \le SizeBloc - 1$$

IDCT : Inverse Discrete Cosinus Transform.

Each block is then copied in the synthesis frame as depicted in figure 3

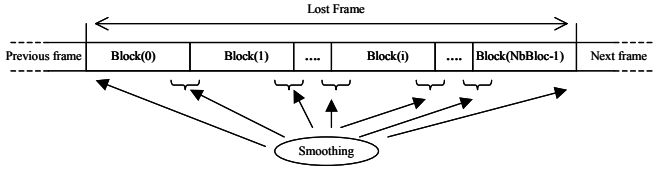

Figure 3: Blocks Concatenation

Smoothing between blocks is realized according to:

$$x(0) = \alpha(0) * x(-1) + (1 - \alpha(0)) * y(0)$$

$x(-1)$: last sample of previous block (or Frame)

y(0) : first sample of current block (or Frame)

$$x(i) = \alpha(j) * x(i-1) + (1 - \alpha(i)) * y(i) \qquad (5)$$

$\alpha(i)$: Smoothing Factor $\alpha(i) = 1 - \frac{i+1}{NbPSmoothing + 1}$

$0 \le i < NbPSmoothing$

In our practical implementation, the value of *NbPSmoothing* is choosen equal to *5* to avoid distortions introduced by long smoothing periods.

In case where *NbBloc*SizeBloc<NbSampleLoss*, missing samples are generated and smoothed using the first sample of last block or missing samples are randomly added using sample interpolation. On the other hand, when *NbBloc*SizeBloc>NbSampleLoss*, extra samples are randomly killed.

Figure 4 and 5 show a concealed frame (30 ms) between 2 voiced frames of a female speech signal.

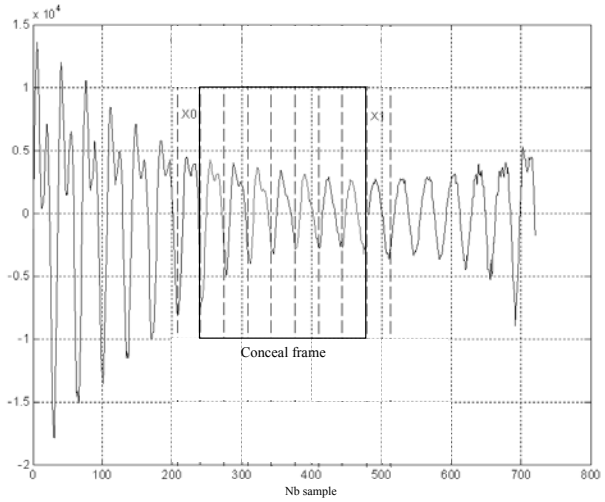

Figure 4: Concealed Signal

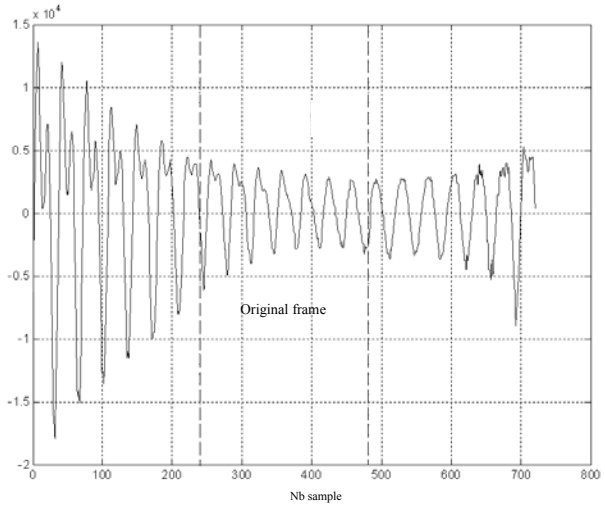

Figure 5: Original Signal

In this case, we can notice that the concealed frame is very close to the original frame. One should notice that the number

of synthesised blocks depends on pitch period. As a consequence, for male speech signal, the number of concealed blocks is generally lower than for a female speech signal.

The main advantage of this technique is the good behaviour for the correction of a noise toward speech or speech toward noise transition. Figure 6 and 7 show the behaviour of the morphing technique during a transition frame (30ms) for male speech signal.

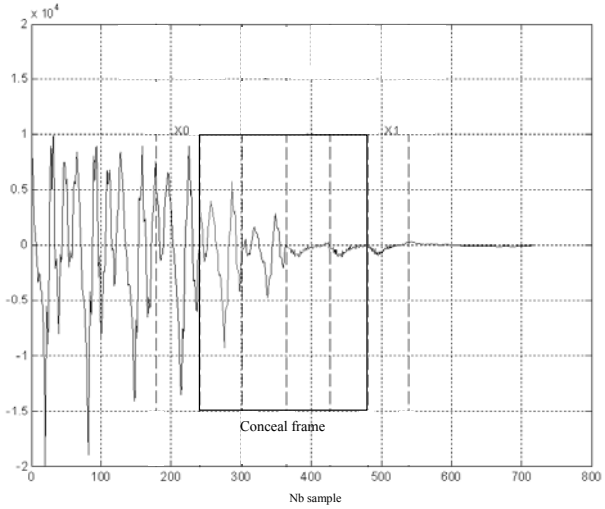

Figure 6 : Speech to noise transition concealment

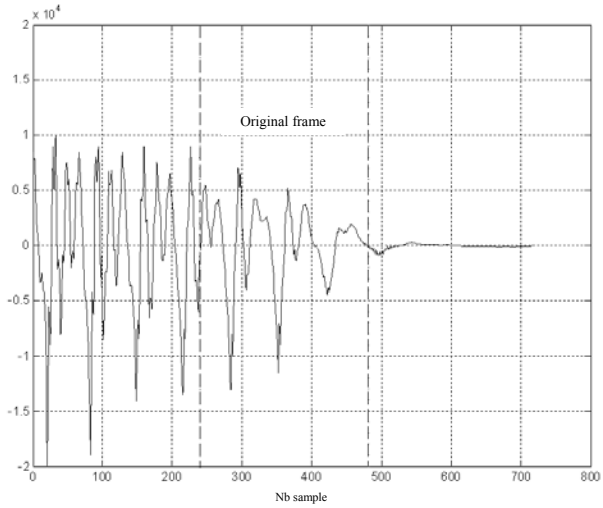

Figure 7 : Original speech to noise transition

Figures 8 and 9 show that the proposed technique can correct with good performances the loss of 2 consecutives frames (60 ms). We can notice that the concealed speech to noise transition is more voiced than original frame. In an enhanced morphing technique the voiced duration could be controlled.

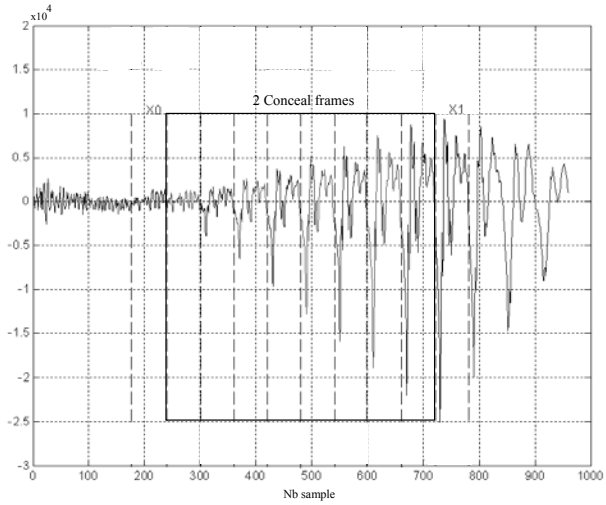

Figure 8 : Two concealed frames

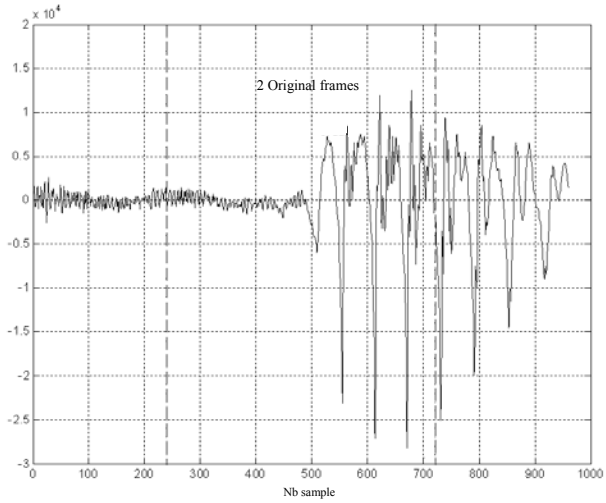

Figure 9 : Original two frames

To further demonstrate the usefulness of the proposed solution, we present in the next section subjective comparison tests with other existing techniques.

3. COMPARISONS AND RESULTS

3.1. Test environment

Ten subjects were participating to an informal test: they were asked to listen to coded speech signals that have been corrected by different concealment techniques. Two speech coders (G.711 and G.723.1) were independently tested. Five concealment techniques were used for both codecs (Previous Frame Copy: PFC, double Sided Periodic Substitution: DSPS, ITU-T recommended technique defined for each specific coder: G.711 or G.723.1, enhanced COMBESCURE technique [5] and Audio Morphing). To evaluate the techniques in various situations of lost frame rate, two series of rate were defined: 5 % and 10 %. The losses can appear by burst, but are usually isolated. The size frame is 30ms. The number of sentences was 15 (8 female and 7 male speech files), thus a total of 150 files were listened to for each coder (G.711 and G.723.1).

For a given *lost frame rate* and each of the 15 files, each subject could listen to each of the 5 concealment techniques as often as wanted in order to choose the technique that provides the best quality. This preferred technique is given a "1" score and the remaining 4 other techniques are given a "0" score. The subjects could choose either 1 or 2 preferred techniques. For example, if a single technique is given for each sentence a "1" score, the final score value for this technique would be "15" that is the maximum possible value *per lost frame rate*. The scores are then averaged over the number of subjects. Results are shown in Figures 10 for G.711 results and Figure 11 for G.723.1 results.

3.2. Results

Speech quality is coder dependant (G.711 speech quality being better than G.723.1 speech quality), it is therefore important to differentiate the results.
The Audio Morphing technique and the enhanced COMBESCURE technique obtain higher scores than the others. However the Audio Morphing technique provides better quality especially for high lost rate where the number of lost bursts is more important than at low lost rate.

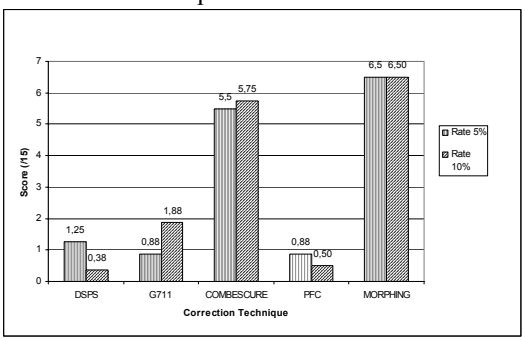

Figure 10: G.711 coder results

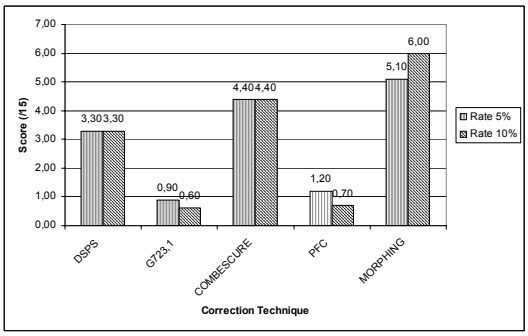

Figure 11: G.723.1 coder results

When the G.711 coder is used (Figure 10), DSPS, recommended ITU-T technique, and PFC do not conceal the missing signal in acceptable way for the tested rates. It can also be noticed that the G.711 technique is adapted to short duration lost period (typically 10ms). In our experiment the packet lost was 30ms long, which penalizes the G.711 technique. For the DSPS and PFC techniques we found that they introduce too much distortion with regards to the global G.711 speech quality, which shows that they are not adapted.

When the G.723.1 coder is used (Figure 11) the ITU-T recommended technique seems to give the same performance as PFC. In fact, G.723.1 coder is a recursive coder, so that erroneous estimation of its state variables degrades the quality of the next synthesized frame. The used of the frame B (frame after the loss frame), even if it is non-fully "coherent" (state variables of the frame A are used to synthesize frame B), improve the quality of the concealment in the case of DSPS and Audio Morphing techniques. However, the enhanced COMBESCURE technique seems to give good performance too.

The authors thank Dominique PASCAL (FT R&D) for her advices concerning the subjective test method.

4. CONCLUSION

We propose in this paper a new concealment technique for lost audio frames. This method uses previous and next frames knowledge, pitch model and morphing principle to synthesize lost frame. Based on subjective results it has been shown that proposed technique improves the quality of the frame correction for strong lost rate (5 % and 10 %). So this technique seems to be particularly suited for VOIP application on Wild Internet.

REFERENCES

[1] ITU Rec., "G.723.1, Dual rate speech coder for multimedia communication transmitting at 5.3kbit/s and 6.3kbit/s," 1996.
[2] D.J. Goodman, G.B. Lockhart, O.J. Wasem, W.C. Wong, "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-34, December 1986, PP. 1440-1448.
[3] AT&T (D.A. Kapilow, R.V. Cox), " A high quality low-complexity algorithm for frame erasure concealment (FEC) with G.711," Delayed Contribution D.249 (WP 3/16), ITU, may 1999.
[4] H. Kobayashi, T. Shimamura, "A weighted autocorrelation method for pitch extraction of noisy speech," Proc. of ICASSP conference, 2000.
[5] P. Combescure et al, "A 16, 24, 32 Kbit/s Wideband Speech Codec Based On ATCELP," Proc. of ICASSP 1999.
[6] J. Tang, "Evaluation of Double Sided Periodic Substitution (DSPS) Method for Recovering Missing Speech in Packet Voice Communications," IEEE Computers and Communications, pp. 454-458, 1991.
[7] RFC 1889, RTP : A Transport Protocol for Real-Time Applications.