AUDIO PACKET LOSS CONCEALMENT USING SPECTRAL MOTION

Seyed Kamran Pedram, Saeed Vaseghi, Bahareh Langari

School of Engineering, Brunel University, London, UK, UB8 3PH
Emails: seyed.pedramrad@brunel.ac.uk, saeed.vaseghi@brunel.ac.uk, bahareh.langari@brunel.ac.uk

ABSTRACT

This paper presents a packet loss concealment (PLC) method with applications to VoIP, audio broadcast and streaming. The problem of modeling of time-varying frequency spectrum in the context of PLC is addressed and a novel solution is proposed for tracking and using the temporal motion of spectral flow. The proposed PLC utilizes a time-frequency motion (TFM) matrix representation of the audio signal where each frequency is tagged with a motion vector estimate. The spectral motion vectors are estimated by cross-correlating the movement of spectral energy within sub-bands across time frames. The missing packets are estimated in TFM domain and inverse transformed to the time-domain. The proposed method is compared with conventional approaches using objective performance evaluation of speech quality (PESQ), and subjective mean opinion scores (MOS) in a range of packet loss from 5% to 20%. The results demonstrate that the proposed algorithm improves performance.

Index Terms—Audio, VoIP, PLC, TFM, Motion compensation

1. INTRODUCTION

The use of the internet, Wi-Fi and mobile devices for transmission and reception of music, speech, broadcast audio and podcasts is pervasive. Audio streaming over the internet involves coding short segments of a digital audio signal which are packed into small data-packets and transmitted. At the receiver the packets are decompressed into waveforms. However, internet providers utilize real-time services, such as VoIP and streaming media based on user datagram protocol/IP (UDP), which for speed reasons unlike transmission control protocol (TCP) do not guarantee quality of service; data packets can be lost, or discarded due to congestion at the routers, network outage or fading of radio signals.

There are two broad approaches for mitigating the degradation in quality due to audio packet loss [1, 2]:

(1) Sender-based packet recovery methods, where the sender changes the encoded bit stream, adding resilience in the form of some redundancy or side information that the receiver can use for the recovery of lost packets. Methods include forward error correction (FEC) and multiple descriptions coding methods [1, 2].

(2) Receiver-based PLC algorithms only use the received packets to estimate the missing signals. Examples are noise substitution, packet repetition, spectral interpolation [2], waveform extrapolation [3], (ITU) waveform substitution (G.711, G.722) [4, 5], model-based extrapolation [6], LP-HNM model of speech [7], pattern matching, overlap-add time-scale modification [8], autoregressive models [9], sinusoidal modeling [10].

Most packet loss methods do not specifically address the important issue of time-variation of the speech spectral parameters on the replacement of lost packets and the solutions that may account for the time-variations. A number of methods that may lend themselves to adaptation for time-varying estimations, such as model- based codecs and overlap-add synthesis, have not been fully investigated in terms of their capacity to provide improved PLC in time varying environments.

The method proposed in this paper is receiver-based PLC operating on the time-frequency signal representation. A novel contribution of this paper is the introduction of time-frequency motion (TFM) matrix and its application to motion-compensated extrapolation/interpolation for audio gap estimation; the idea is similar to that employed in motion-compensated image processing, however, here it is the motion of frequencies across time frames that are estimated and factored in the estimation process.

The remainder of the paper is organized as follows. Section 2, describes time-frequency motion matrix. Section 3 presents the concept of spectral motion vector model and its method of estimation. Simulation of the packet loss in TFM representation is presented in section 4. Section 5 provides comparative evaluation results and section 6 concludes the paper.

2. TIME-FREQUENCY MOTION MATRIX MODEL

The algorithm for TFM matrix representation of audio signals is shown in the flow chart 1. The audio input stream $x(m)$ is segmented into overlapping windowed segments. Successive segments are transformed by a discrete Fourier transform (DFT) and stacked to form a time-frequency matrix given by

$$X_{DFT}(k,l) = \sum_{m=0}^{N-1} x(l*s+m)e^{-j2\pi mk/N} \quad (1)$$

Where N is the segment length , m is the discrete-time, $k = 0, \dots, N-1$ is the discrete-frequency, l is the frame index and s is the segment overlap. Alternatively, a discrete cosine transform (DCT) TFM can be formed as

$$X_{DCT}(k,l) = w(k)\sum_{m=0}^{N-1} x(l*s+m)\cos[\pi(m+0.5)k/N] \quad (2)$$

Where $w(0) = 1/\sqrt{N}$ and $w(1{:}N-1) = \sqrt{2/N}$. For time-varying signals the spectral values $X(k,l)$ can be augmented by the spectral motion variables $\mathcal{M}(k,l)$ to form a state-space Kalman formulation [11], as

$$\begin{bmatrix} X(k,l) \\ \mathcal{M}(k,l) \end{bmatrix} = A\begin{bmatrix} X(k-\mathcal{M}(k,l-1),l-1) \\ \mathcal{M}(k,l-1) \end{bmatrix} + \begin{bmatrix} w(k,l) \\ \dot{w}(k,l) \end{bmatrix} \quad (3)$$

Where A is the state transition matrix and the vector $[w\ \dot{w}]$ is a random input.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

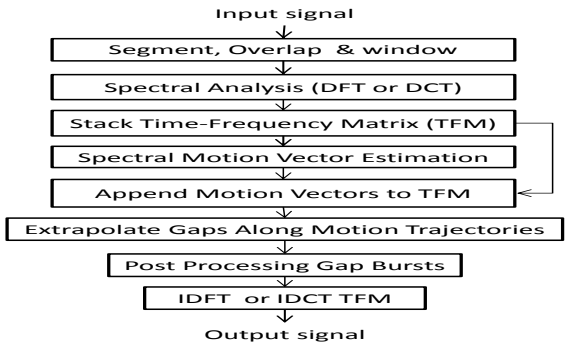

Fig. 1. Time-frequency motion modeling of audio signal

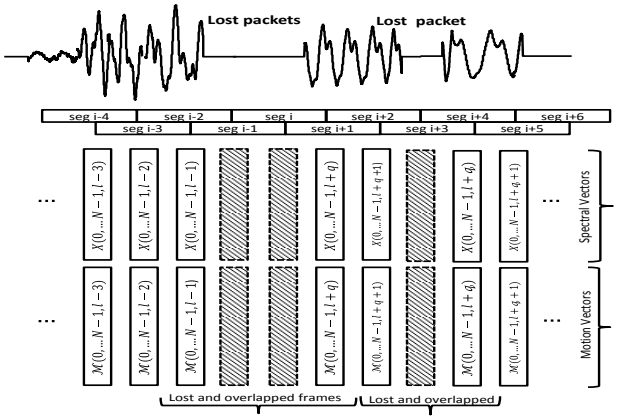

Fig. 2. Transformation of a time-domain signal with gaps into TFMs with spectral motion vectors appended.

Solution of Equation (3) presents a non-trivial problem. A practical solution is shown in Figure 1 where the input signal $x(m)$ is transformed into an augmented spectrogram matrix $X(k,l)$ appended with the spectral motion matrix $\mathcal{M}(k,l)$ as

$$x(m) \overset{FTM}{\Longleftrightarrow} \{X(k,l), \mathcal{M}(k,l)\} \tag{4}$$

Where $\mathcal{M}(k,l)$ describes the motion of the k^{th} frequency. Figure 2, shows a time domain audio signal with missing gaps together with the TFM representation of the signal. Note that a gap of N missing samples is transformed into one or several frequency column vectors in the TFM matrix. Each frequency component of lost packets is estimated from the previous available packets alone or plus future frames.

3. SPECTRAL MOTION VECTOR (SMV) MODEL

In this section a method is proposed for deriving a set of motion vectors $\mathcal{M}_{k,l}$, that when appended to a spectral vector $X(k,l)$ indicates the direction along which the k^{th}frequency of the l^{th} frame moves relative to the previous frame $l-1$ .The proposed method divides the signal spectrum $X(k,l)$ into B overlapping subbands $X_{SB}(k,l)$, $SB = 0,\dots,B-1$. A computational method can be used to estimate the spectral motion in each subband between the frames $l-1$ and l as

$$\mathcal{M}_{k,l} = \underset{i}{argmax}\left(\sum_{i\in X_{SB}} f\big(X_{SB}(i,l), X_{SB}(i,l-1)\big)\right) \tag{5}$$

Where the function $f(.)$ compares two successive spectral frames in order to caclulate a spectral motion vector. The computational method used here is the cross correlation (CC) of the frequency bands across two sucessive time frames; the motion of the k^{th} subband of the l^{th} frame relative to the $(l-1)^{th}$ frame, $\mathcal{M}_{k,l}$, is determined from the position of the CC lag corresponding to the peak of the CC function as

$$\mathcal{M}_{k,l} = \underset{i}{argmax}\left(\sum_{i\in X_{SB}} X_{SB}(i,l)\, X_{SB}(i,l-1)\right) \tag{6}$$

Where X_{SB} is the subband in which the frequency k resides. As an alternative method we experimented with the spectal motion vector estimated by minimising the mean squared error distance between the spectral vectors.

Since the spectral motion over time would be quantised to the frequency resolution $Df = Fs/N$, the technique of zero-padded DFT can be used to yield a higher resolution interpolated spectrum and hence obtain a finer quantisation of the spectral motion variable $\mathcal{M}_{k,l}$. The estimates of motion vector may be smoothed over time using a first order recursive equations as

$$\widehat{\mathcal{M}}_{k,l} = a\widehat{\mathcal{M}}_{k,l-1} + (1-a)\mathcal{M}_{k,l} \tag{7}$$

Where $\widehat{\mathcal{M}}_{k,l}$ denotes the smoothed motion vector and the variable a may be set to a value in the range 0.95-0.99.

4. EXTRAPOLATION/INTERPOLATION IN TIME-FREQUENCY MOTION (TFM) MATRIX

It would be incorrect to extrapolate a missing frequency partial $X(k,l)$ from the same frequency bin, k, of the previous frames, $l-i$, $X(k,l-i)\ i = 1,2\ \cdots$, if, as shown in Figure 3, over the time, the spectral power is moving across the frequency bins. The correct method is to estimate the motion trajectory of the frequency partials and to extrapolate the k^{th} frequency partial from $X\big(k-\mathcal{M}_{k,l-1}, l-1\big), X\big(k-\mathcal{M}_{k,l-2}, l-2\big), \dots$ where the variable $\mathcal{M}_{k,l-1}$ indicates the relative motion of the k^{th} frequency partial between the frames $l-1$ and l. In this work a moving average motion-compensated method is used for extrapolation. Assuming that the l^{th} packet is lost, and needs to be extrapolated using the previous Q packets, the spectral motion-compensated extrapolation formula can be expressed as

$$X_{extrap}(k,l) = \sum_{j=1}^{Q} c_j X\big(k-\mathcal{M}_{k,l-j}, l-j\big) \tag{8}$$

Where c_j are the coefficients of an extrapolation polynomial of order Q. The polynomial coefficients may be calculated and updated from a least square error fit of an n^{th} order polynomial such as a linear or second order or spline function to the frequency tracks immediately preceding the missing gaps [7].

Since the frequency partials and motion of the missing frame l are unavailable, there is a question as to which frequency track, k', from the previous frames $l-j$ should be extrapolated into the k^{th} partial of the missing frame. The frequency track k' from the past frame,$l-j$, passing closest to the k^{th} frequency partial of the missing frame l can be estimated, using a first order estimation as

$$k' = \min_{i=1:N}\big(abs(i + j\mathcal{M}_{k,l-j} - k)\big) \tag{9}$$

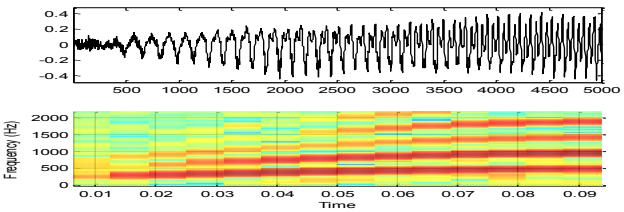

Fig. 3. Illustration of a speech segment where the spectral power is moving across the frequency over the time frames

Where i are the frequency partials of frame $l-j$, j is the distance, in terms of the number of frames, of the missing frame from the closest available frame. Equation (9) can be modified to a spectral motion-compensated interpolation method, using Q past and P future spectral vectors as

$$X_{Interp}(k,l) = \sum_{j=1}^{Q} b_j X\left(k - \mathcal{M}_{k,l-j}, l-j\right) + \sum_{j=1}^{P} b_{Q+j} X\left(k - \mathcal{M}_{k,l+j}, +j\right) \qquad (10)$$

Where the coefficient vector $[b_1, \dots\ b_Q, b_{Q+1}, \dots\ b_{Q+P}]$ operate on the available past and future frames $[l-1, \dots\ l-Q, l+1, \dots, l+P]$. Interpolation would be useful in cases where a delay of one or two frames is not critical and when the future frames have not been lost. In this work for extrapolation of each frequency frame, a linear extrapolator, passing through the past $Q = 3$ frames, is used. For interpolation we also included $P = 2$ future frames. These rules can be extended for burst losses and reconstructed in the same manner.

5. PERFORMANCE EVALUATION

The proposed PLC method is evaluated for both objective and subjective tests on audio speech with bandwidths of 10 kHz to 20 kHz and the corresponding sampling rates of 16 kHz to 48 kHz which were resampled to 8 kHz or 16 kHz. The segment length was set to 25ms and segment shift to a quarter the segment length. Each sample is represented by 16 bits.

The DFT or the DCT are used as alternatives for forming TFMs. To mitigate end-discontinuity effect, the widely used Hanning window is applied to signal segments. In addition, as shown in Figure 2, for the purpose of forming TFM, successive speech segments are overlapped. For half overlap, when a packet loss occurs, only one previous TFM frame, and for 75% overlap, three previous TFM frames which overlap with the lost one are unavailable for extrapolation. This rule is applied for interpolation method as well.

The packet loss model that used here is a binary-state independent identically distributed (IID) Bernoulli model with packet loss rates varying from 5% to 20% [7].

To mitigate unnatural sounds that inevitably result from extrapolation of long sequence of missing packets, an attenuation technique is applied which is common and similar to other algorithms such as ITU standard G.711 [4].

The results are compared to some alternative methods that estimated the packet loss such as i) a LP-HNM model of speech where the spectral envelope is modeled using a LSF representation of a linear prediction (LP) model [7], ii) method proposed in [10] which is based on interpolation of harmonics in a sinusoidal

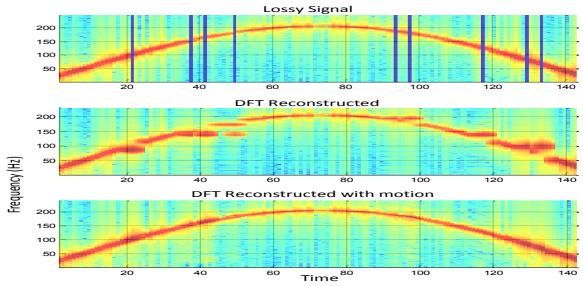

Fig. 4. Spectra of simulated time-varying signal, from top: lossy signal, DFT-without motion and DFT-TFM (with motion) signal.

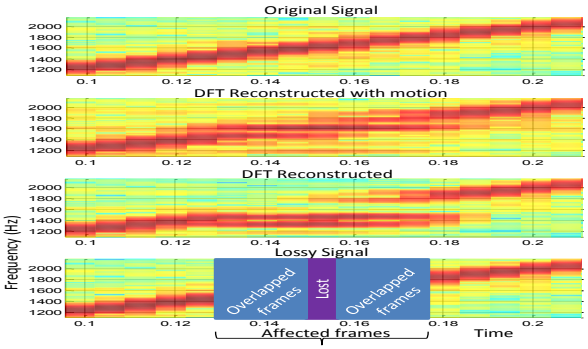

Fig. 5. Spectra of simulated signal, from top: original signal, DFT-TFM , DFT, and signal with 20% Bernoulli frame loss.

model (SM) for excitation, iii) The Multirate technique which is based on time-domain AR modeling of the signals[9], and iv) the ITU standard packet loss concealment algorithm G.711 [4].

Due to space limitations, the results presented in this paper relate to DFT extrapolation only. Further improvements obtained from DCT and interpolation will be presented in a future paper.

5.1 Experiments on Simulated Time-Varying Signals

The proposed method (DFT-TFM) firstly is tested and applied to simulated time-varying signals and compared with DFT method without motion as shown in Figure 4 and Figure 5. Visibly the proposed algorithm improved performance especially when the lost occurred in different frequency bins. In addition, Figure 5 provides a comparison of a part of the spectra of an original simulated signal with lost packets and the reconstructed spectra with proposed technique and DFT model without motion.

For both experiments, the packet loss rate is set to 20% Bernoulli frame loss. The overlap between successive signal windows is set to 75%. Hence due to overlap, three previous and future frames are affected and cannot be used to estimate the lost frames. It is clear that the frequency bins of the spectral power change across time frames and extrapolations from the previous frequency frames couldn't give the correct result while the motion trajectory follows the movement for estimating the lost packets.

5.2 Objective Evaluation

In this section, signals are reconstructed by the proposed method and the conventional methods mentioned earlier. Perceptual Evaluation of Speech Quality, (PESQ, ITU-T P.862) [12], is utilized for comparison. The results are calculated and averaged for 50 sentences randomly selected from the TIMIT database, employing the Bernoulli distributed random lost frames [7].

Table 1. Performance of different algorithms for restoration of Bernoulli generated gaps.

	PESQ Result			
Loss Rate %	5	10	15	20
Av.Gap Length	1.05	1.11	1.18	1.25
DFTM	3.91	3.28	3.10	2.92
LP-HNM [7]	3.82	3.20	3.02	2.81
G.711 [4]	3.80	3.19	2.97	2.73
SM [10]	3.60	2.98	2.71	2.56
Multirate [9]	3.58	2.75	2.54	2.35
Distorted	3.43	2.71	2.49	2.28

The results in Table1 indicate that the proposed method performs better and achieved higher score than other algorithms in terms of PESQ measurement and shows that the performance of G.711 and LP-HNM models are good at lower loss rates, but deterioration is increased for higher loss rates. The LP-HNM and SM algorithms modeled-interpolated the envelopes throughout the gap and perform better than G.711 and Multirate algorithm at higher loss rates.

Furthermore, the proposed technique illustrations more robustness to longer frame loss compared to other methods. It should be consider that the G.711 PLC method is generally designed to cope with 60 ms of loss and it deteriorates quickly for losses longer than that [4, 7].

5.3 Subjective Evaluation

Five random wideband speech samples used in this experiment are listed in Table 2. Signals are about 14-18 seconds long, with fixed packet loss rate of 20% and average gap lengths of 2, 5, 7 frames utilizing Bernoulli distributed random lost frames [7]. To evaluate the result after applying the gaps, each sample was reconstructed using the proposed technique (DFT- TFM), LP-HNM [7] and G.711 [4] algorithms. Subjective test were carried by 12 student listeners aged between 22 to 31 years in a quiet room using headphones. The students were asked to compare speech degraded by packet losses with enhanced speech where the lost packets were replaced using DFT- TFM, LP-HNM and G.711 algorithms. Listeners' preferences were recorded by using the mean opinion score (MOS) point base assessment where the scores vary from 1 (bad) to 5 (imperceptible degradation).

Table 3 displays the result of this analysis which the corresponding confidence intervals were ± 0.12. The result verifies that the proposed technique compared to other methods perform better and achieve higher quality.

Table 2. Audio speeches and music.

No.	File Name	Type of Audio	Sampling Rate
1	Barack Obama - Berlin	Speech	22050
2	Speech on Womens Right toVote	Speech	44100
3	Civil Rights- United States	Speech	44100
4	Abraham Lincoln	Speech	44100
5	Adele - Take It All	Music	48000

Table 3. Comparative subjective results of proposed methods with a loss rate of 20% against LP-HNM and G.711.

Restoration Method	DFT-TFM			LP-HNM [7]			G.711 [4]		
Av.Gap Length	2	5	7	2	5	7	2	5	7
Subjective Score	4.10	3.41	2.97	3.93	3.29	2.88	3.98	3.25	2.72

6. CONCLUSION

In this paper, the problem of restoration of gaps in audio signal was addressed and a novel solution of packet loss concealment is presented for audio signals based on time-frequency motion (TFM) using discrete Fourier transform (DFT). The novel aspect of this methodology is the introduction of TFM and its application to motion-compensated extrapolation and interpolation for audio. The spectral motion vectors were estimated by dividing the signal bandwidth into several overlapping sub-bands. The cross correlation (CC) of the frequency bands across time frames has been used for motion estimation. The objective and subjective evaluation experiment demonstrates that the proposed technique compares well with conventional methods with the results in superior output quality in terms of PESQ and MOS scores. In addition, comparison between DFT and discrete cosine transform (DCT) as well as discussion about further improvements which obtained from interpolation technique will be presented in a future paper.

REFERENCES

[1] Xiguang Zheng; Ritz, C., "Hybrid FEC and MDC models for low-delay packet-loss recovery," Signal Processing and Communication Systems (ICSPCS), 2011 5th International Conference on, pp.1,6, 12-14 Dec. 2011

[2] H. Ofir, "Packet loss concealment for audio streaming" Master's thesis, Technion-Israel Inst., Haifa, Israel, 2006.

[3] Juin-Hwey Chen, "Packet loss concealment based on extrapolation of speech waveform," Acoustics, Speech and Signal Processing. ICASSP 2009. IEEE International Conference on, pp.4129,4132, 19-24 April 2009

[4] Appendix I: A High Quality Low-Complexity Algorithm for Packet Loss Concealment With G.711 ITU-T Recommend. G.711, Sep. 1999.

[5] J. Thyssen, R. Zopf, J.-H. Chen, and N. Shetty, "A Candidate for the ITU-T G.722 packet loss concealment standard," Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, vol. 4, pp. IV-549 – IV-552, April 2007.

[6] J.-H. Chen, "Packet loss concealment for predictive speech coding based on extrapolation of speech waveform," ACSSC 2007 Conference on Signal, Systems and Computers, California, USA, pp. 2088-2092, Nov. 2007.

[7] Zavarehei, E.; Vaseghi, S., "Interpolation of Lost Speech Segments Using LP-HNM Model With Codebook Post-Processing," Multimedia, IEEE Transactions on , vol.10, no.3, pp.493,502, April 2008

[8] Merazka F. Packet loss concealment using time scale modification for CELP based coders in packet network. In: 40th southeastern symposium on system theory. pp. 84–7, 6–18 March; 2008.

[9] P. A. A. Esquef and L. W. P. Biscainho, "An efficient model-based multirate method for reconstruction of audio signals across long gaps," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1391–1400, Jul. 2006.

[10] J. Lindblom and P. Hedelin, "Packet loss concealment based on sinusoidal modeling," in Proc. IEEE Workshop on Speech Coding, Ibaraki, Japan, pp. 65–67, Oct. 2002.

[11] Q. Yan, S. Vaseghi, E. Zavarehei, et al, Kalman tracking of linear predictor and harmonic noise models for noisy speech enhancement, Computer Speech & Language, vol. 22, no. 1, pp. 69-83, Jan. 2008.

[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in Proc. Acoustics, Speech, and Signal Processing, ICASSP, vol. 2, pp. 749–752, May 2001.