RECONSTRUCTION OF MISSING SPEECH FRAMES USING SUB-BAND EXCITATION

Kai Clüver, Peter Noll

Institut für Fernmeldetechnik, Technische Universität Berlin
Einsteinufer 25, D-10587 Berlin, Germany

ABSTRACT

A new reconstruction method for frame erasures in speech transmission is presented which is based on parameterization of the speech signal by means of linear prediction (LPC) and voicing analysis. The problem of generating partially voiced substitute speech signals is solved by performing separate voicing decisions in sub-bands. The method yields considerable improvements compared with silence substitution for frame erasure ratios of up to 10 % or even 20 %. The combination of the reconstruction method with adaptive speech coders showed virtually the same good results for forward adaptation, whereas a higher degradation is caused by backward-adaptive coders.

1. INTRODUCTION

In communication networks such as packet and ATM networks, or wireless systems such as land mobile networks and future Personal Communication Networks (PCN), the digital signals are transmitted in uniform frames. The frame format results in packet loss caused by network congestion or excessive delay of some packets, or, equivalently, frame erasures due to fading in the mobile transmission link.

The main problem in dealing with frame erasures in speech signals is to generate a substitute for the missing speech frame which minimizes the perceptibility of the erasure. Methods of either interpolation or extrapolation have been investigated; as the signal delay in a two-way telephone link is critical for unimpaired flow of the conversation, only extrapolation of the speech signal preceding the missing frame is usually considered. In previous work on PCM packet transmission, repetition at the pitch period showed the best results for direct substitution of the speech waveform [1]. More recently, further improvement could be obtained by substitution of the linear prediction (LPC) residual instead of the speech signal [2]. The reconstruction of frame erasures in sub-bands has also been investigated [3].

In adaptive speech coders, frame erasures cause the state variables and buffer contents in the decoder to differ from those in the encoder. Due to slow convergence after the frame erasure, additional distortions in the speech signal may occur, especially with backward-adaptive coders. In recent work [4, 5], the effects of frame erasures on the backward-adaptive LD-CELP Coder were investigated.

In this work, the generation of the substitute speech signal is considered, assuming PCM coding at first, i. e. ignoring adaptation problems. A new reconstruction method is presented and its performance is discussed. Finally, some experiments with adaptive speech coders are reported.

2. RECONSTRUCTION OF MISSING FRAMES

2.1 Parameterization of the Speech Signal

A simple way of extrapolating the speech signal during a frame erasure is the repetition of the preceding frame. Only small improvements compared with silence substitution (which is regarded here as the reference) can be achieved, as voiced speech will show discontinuities at the edges of the reconstructed frame. Better results are obtained by analysing the speech signal immediately preceding the missing frame and retaining important speech parameters while generating the substitute signal. This is done in [1], where the pitch period is determined and the preceding signal is repeated at this period during the reconstruction of the missing frame, thus retaining the fundamental frequency. Further parameterization can be accomplished by spectral analysis. With an additional LPC analysis, subsequent pitch replication of the residual signal, and LPC synthesis, both pitch and spectral envelope are retained during the reconstruction of the missing frame [2].

While this latter method yields good results for frames of up to about 10 ms duration, the speech signal shows annoying artifacts with longer frames, as periodical reconstruction (which is inappropriate for unvoiced speech) now becomes perceptible. As a consequence, voicing analysis should be added. Experiments with a binary voiced/un-

voiced decision did not result in any improvement, as partially voiced speech cannot adequately be reconstructed. Consequently, a frame erasure recovery method for longer frames should include a mixed excitation of the LPC synthesis filter.

2.2 The Sub-Band Reconstruction Method

The new method presented here solves the problem of reconstructing partially voiced speech by performing separate voicing decisions in sub-bands. The structure is depicted in fig. 1. For each valid frame, LPC coefficients for a 12th order linear predictor are calculated from the decoded speech samples. The decoded speech is passed through an adaptive prediction error filter and the resulting LPC residual is then split into eight sub-bands. The filter bank consists of Mth-band filters [6] which allow a perfect reconstruction of the full-band signal. The sub-bands are added together to form the excitation signal of the synthesis filter which, like the prediction error filter, is adapted using the LPC coefficients. As the filter bank introduces some signal delay, however, the synthesis filter update is delayed with respect to the update of the prediction error filter. Thus, while the decoder receives valid frames, the synthesis filter output is identical to a delayed version of the decoded speech signal.

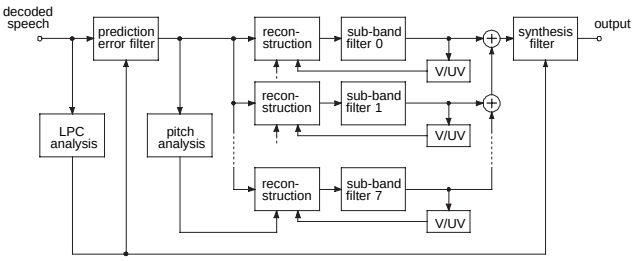

Fig. 1 Block diagram of the missing frame reconstruction method

During frame erasures, the synthesis filter coefficients of the last valid frame are retained. From the preceding LPC residual, the pitch period is determined by means of autocorrelation analysis. The sub-bands are classified as either voiced or unvoiced by analysing the preceding output signals of the filter bank. Then, the input signals of the filters are reconstructed separately for each sub-band. For voiced bands, the previous filter input is periodically repeated, the pitch period being the same for all voiced sub-bands. For unvoiced bands, an appropriately scaled white noise signal is used. The sum of the filter output signals now forms the substitute excitation signal for the synthesis filter. Possible discontinuities at the boundary of the substitute excitation and the residual from the succeeding frame are smoothed by merging both signals at the beginning of the next valid frame.

3. DISCUSSION

The proposed sub-band reconstruction method was investigated for frames of 20 ms duration, which is common for mobile networks such as GSM. Memoryless PCM speech coding was assumed, i. e. error-free decoding of the first valid frame following frame erasures. For voiced speech, most sub-bands (if not all) are classified voiced, which results in a periodical repetition of the previous excitation. Fig. 2(a) shows a 60 ms segment of voiced speech; the result of substituting silence (i. e. zero-valued samples) for an erased frame is shown in fig. 2(b). The sub-band reconstruction of the LPC residual results in the waveform of fig. 2(c) which differs only slightly from the original speech waveform, since voicing, pitch, and spectral envelope are kept at constant values.

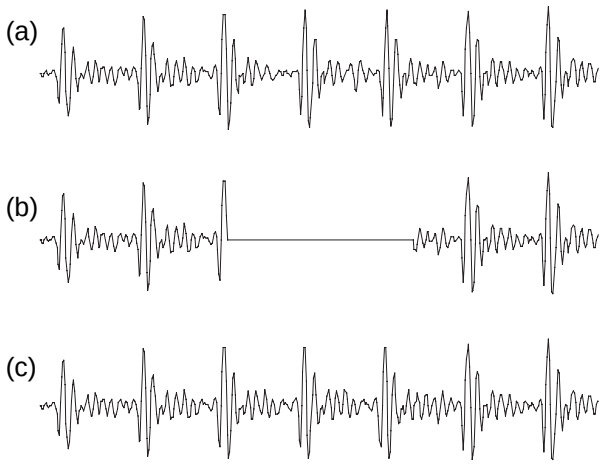

Fig. 2 Waveforms of voiced speech: (a) without frame erasure, (b) with silence substitution, (c) with sub-band reconstruction of the LPC residual

An example for the reconstruction of a frame erasure during a voicing transition is depicted in fig. 3 (the waveforms (a), (b), (c) correspond to those in fig. 2). Only a few sub-bands are voiced, and the reconstructed excitation signal of the synthesis filter is composed of sub-band signals which are either periodical or random. Although

the true beginning of the unvoiced speech is lost, the reconstruction method renders the frame erasure almost imperceptible.

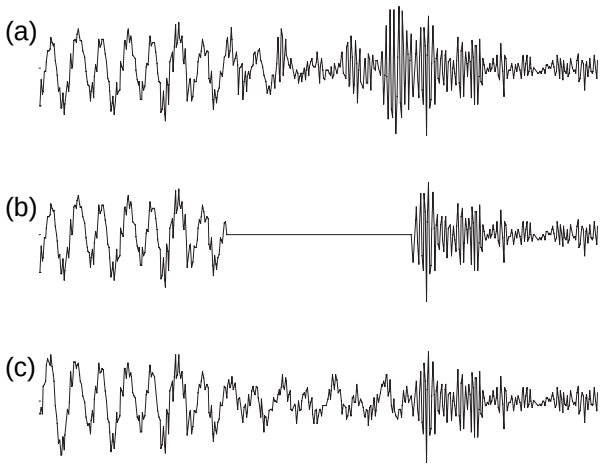

Fig. 3 Speech waveforms of a voicing transition: (a) without frame erasure, (b) with silence substitution, (c) with sub-band reconstruction of the LPC residual

FER	silence	sub-band
0 %	3.97	
1 %	3.42	4.00
5 %	2.11	3.36
10 %	1.44	2.97
20 %	1.22	2.28

Table 1 MOS results for silence substitution and sub-band reconstruction of the LPC residual

The quality of the method was assessed in a mean opinion score (MOS) listening test; table 1 shows the results for a frame size of 20 ms, together with the MOS ratings for silence substitution. At a frame erasure ratio (FER) of 1 %, no MOS degradation was found with the sub-band reconstruction method. At 10 % FER, fair quality (i. e. MOS of 3) is achieved; this corresponds to a gain of more than 1.5 compared with silence substitution. Even at an FER as high as 20 %, the MOS gain is still around 1. The 95 % confidence interval for this test was between 0.2 and 0.3, so that differences of more than about 0.5 can be regarded as statistically significant.

4. LOST FRAME RECOVERY FOR ADAPTIVE SPEECH CODERS

While developing the reconstruction method it was assumed that the speech signal succeeding a frame erasure can be decoded without error. This is, however, only true for non-adaptive, i. e. memoryless, PCM coders. The state variables of adaptive decoders may need some time to recover from frame erasures, which may result in additional degradation of the reconstructed speech signal. On the other hand, possibly identical structures of the decoder and the reconstruction method can be merged in order to reduce the computational complexity. The reconstruction method of section 2.2 was tested with two adaptive coders, forward-adaptive 6.5 kbit/s CELP and G.726 backward-adaptive 32 kbit/s ADPCM.

4.1 Forward-Adaptive CELP

The frame size of the CELP coder [7] is 20 ms; the synthesis filters are updated once per frame. The excitation parameters are updated every 5 ms, i. e. four times per frame. The reconstruction of missing frames is carried out in the decoder, exploiting part of its structure: using the decoded parameters of the previous frame, neither LPC nor pitch analysis need be done. During missing frames, the synthesis filter (and post-filter) coefficients are retained, and the most recent adaptive codebook index serves as pitch period. Only the filter bank and the sub-band voicing analysis have to be added to the decoder. The excitation signal of the synthesis filter is then reconstructed in sub-bands according to section 2.2.

The substitute excitation is also used to update the adaptive codebook. This posed no problems for single frame erasures. In the case of bursty frame erasures, however, the contents of the adaptive codebooks in the encoder and decoder differ considerably. In order to avoid spurious pulses in the first valid frame of voiced speech, the adaptive codebook in the decoder is reset after two or more consecutive missing frames.

The performance of the reconstruction method with the CELP decoder was checked in informal listening tests. For random frame erasures, no additional degradation (compared to PCM) was found in the reconstructed speech. For bursty erasures, some variation of loudness and voicing is introduced on account of the adaptive codebook reset, but the degradation remains small.

4.2 Backward-Adaptive ADPCM

The state variables of the G.726 ADPCM coder are updated with every sample of valid frames. The decoder

output signal forms the input of the structure shown in fig. 1, so that all anlysis and filtering is performed on the decoded speech signal. During missing frames, the decoder variables are frozen, while the reconstruction is done outside the ADPCM decoder, again in sub-bands of a synthesis filter excitation signal.

When sample-by-sample updating of the variables is resumed after a frame erasure, the adaptation speed of the excitation gain in the decoder may be higher than in the encoder. This results in annoying effects in the output signal due to extremely high decoder gain values. The problem is solved by setting the adaptation speed parameter to its minimum value and only slowly "thawing" it subsequently. This, however, leads to a slower convergence of the decoder after frame erasures. Alternatively, the adaptation speed of both encoder and decoder could be set to its maximum value, thus giving up the bit-stream compatibility with G.726 for faster convergence of the decoder. Even then, however, occasional clicks will be present in the output speech signal, again requiring a slowing down of the decoder gain adaptation after frame erasures.

Informal listening tests showed that while the reconstruction method with G.726 ADPCM yields satisfactory results for lower frame erasure ratios, the speech quality at 10 % FER is inferior to the quality of PCM or forward-adaptive CELP. The frame erasures lead to a considerable variation of the output level; additionally, a rough, less voiced quality is introduced into the speech signal. The reason for this is the loss of speech transitions during erased frames: the time between two frame erasures may then be insufficient for a complete convergence of the decoder variables.

5. CONCLUSION

A new method for the reconstruction of missing speech frames was presented which is based on LPC and pitch analysis together with voicing decisions in sub-bands. For a frame size of 20 ms, high speech quality gains compared with silence substitution are achieved. At frame erasure ratios (FER) of up to 10 %, the method yields a MOS of 3 or higher. Combined with a forward-adaptive CELP coder, virtually the same good results are obtained, whereas satisfactory results can only be achieved at lower FERs with backward-adaptive G.726 ADPCM.

The main motive for employing backward-adaptive speech coders is a low signal delay. The assembly of speech frames, on the other hand, requires buffering and, consequently, delaying the speech signal. Thus, the delay advantage of backward adaptation is lost, while a quick recovery from frame erasures is difficult to achieve. In order to ensure the least possible quality degradation under frame erasures, forward-adaptive speech coders should preferably be employed for frame-based transmission networks.

ACKNOWLEDGEMENTS

The authors wish to thank H. Klaus and E. Fuhrmann for their help with the MOS listening test.

This work was supported by the DFG (Deutsche Forschungsgemeinschaft), Bonn, Germany.

REFERENCES

[1] O. J. Wasem, D. J. Goodman, C. A. Dvorak, H. G. Page: The Effect of Waveform Substitution on the Quality of PCM Packet Communications, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 36, no. 3, March 1988, pp. 342-348

[2] K. Clüver: An ATM Speech Codec with Improved Reconstruction of Lost Cells, to be presented at EUSIPCO-96, Trieste, Italy, September 1996

[3] W. C. Wong, N. Seshadri, C.-E. W. Sundberg: Estimation of Unreliable Packets in Subband Coding of Speech, IEE Proceedings-I, vol. 138, no. 1, February 1991, pp. 43-49

[4] C. R. Watkins, J.-H. Chen: Improving 16 kb/s G.728 LD-CELP Speech Coder for Frame Erasure Channels, Proc. ICASSP 1995, pp. 241-244

[5] A. Husain, V. Cuperman: Reconstruction of Missing Packets for CELP-Based Speech Coders, Proc. ICASSP 1995, pp. 245-248

[6] P. P. Vaidyanathan: Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial, Proceedings of the IEEE, vol. 78, no. 1, January 1990, pp. 56-93

[7] K. Clüver, T. Gries, H. Li, P. Noll: Real-Time Implementation of a CELP Codec with Unequal Error Protection, Proc. EUSIPCO 1992, pp. 1541-1544