ANALYSIS AND SYNTHESIS METHOD FOR PACKET SPEECH ENHANCEMENT

S N Koh and P. Sivaprakasapillai
School of Electrical and Electronic Engineering
Nanyang Technological Institute, Singapore 2263

ABSTRACT: This paper describes a novel system which employs the filter bank analysis and synthesis method for the packetisation of speech for transmission in packet data communications. Computer simulations of the system indicate that significant improvement in the perceptual quality of the recovered speech can be obtained even with zero substitutions compared to the conventional technique of straight packetization of PCM speech. Further improvement is possible through frequency-domain component replications.

INTRODUCTION

Recent developments in digital speech transmission suggest that transmission of speech through packet communication networks will become more important in the future. It is therefore important that problems associated with such networks be addressed. The main problem with packet speech transmission is that not all the packets will arrive at the addressee in time and intact. When the packet data network is experiencing excessive traffic, packets of data, which might include speech data, might be dropped at the switching nodes or even at the entry ends after a certain delay interval.

Missing packets due to channel congestion and transmission impairment are the main source of speech degradation in these systems. Missing packets result in gaps of silence intervals in the received speech waveform if no enhancement through post-processing is carried out. With 10 percent missing packets, the perceptual effect of the recovered speech varries from crackling, popping to gargling depending upon the packet size [1].

One way to alleviate this problem partially is through waveform replication technique suggested by Goodman [1,2]. Instead of applying time-domain waveform replication methods, a frequency-domain analysis and synthesis algorithm is investigated for the enhancement of packet speech in this paper.

THE SYSTEM

The most straight forward configuration for sending digitized speech through a packet data network is shown in Figure 1. Analogue speech is first converted into 8-bit samples using log-PCM or 12-bit samples using linear A/D conversion. They are then buffered into 8, 16 or 32 msec packets and appended with appropriate addresses and network control characters for transmission. At the receiver, packets arriving at different times are reordered and assembled for recovery by the D/A conversion.

What is considered in this paper is the application of filter bank analysis procedure implemented through digital signal processor as

a preprocessor as shown in Figure 2. An inverse operation, ie filter bank synthesis, is carried out at the receiver to recover the transmitted speech.

ANALYSIS AND SYNTHESIS METHOD OF ENHANCEMENT

Filter-bank Analysis:

Speech samples x(n)'s are first passed through the analysis filter bank as shown in Figure 3. $H_o(e^{jw})$ is a lowpass analysis filter and $H_i(e^{jw})$ (i = 1,...., N-1) is the ith bandpass filter with cut-off frequencies of (2i-1) fs/2N and (2i+1) fs/2N where fs is the sampling frequency. The output of each filter is translated to baseband by multiplying by $e^{-jw_k n}$ for each k to give $Y_n(k)$. $Y_n(k)$ can be down-sampled by a factor of R as long as $R \leq N$ where N is the number of subbands of the system. In our system R is set to N. The down-sampled $Y_{rR}(k)$'s are then quantized to 12 bits resolution and organised into 64-sample (ie 8 msecs) packets for transmission.

Filer-Bank Synthesis:

To recover the original speech samples $\hat{x}(N)$, the received $\hat{Y}_{rR}(k)$ samples (quantized values of $Y_{rR}(k)$'s) are unsampled by inserting N-1 zeros in between every 2 $\hat{Y}_{rR}(k)$ samples and then lowpass filtered by $H_o(e^{jw})$. The filtered sequences are then translated to their original frequency positions by multiplying by $e^{jw_k n}$ for each k to give $\hat{Y}_n(k)$'s which are then summed together to form $\hat{x}(n)$, ie

$$\hat{x}(n) = \sum_{k=o}^{N-1} \hat{Y}_n(k) e^{jw_k n} \qquad (1)$$

Advantage of Filter-Bank Method for Packet Speech Transmission:

The advantage of the systeem described so far is that when there are missing packets, the recovered speech would not exhibit silence gaps as happens in straight packetization of PCM speech. Figure (4.c) shows the effect on the recovered speech with three 64-sample packets missing. Evidently, the three packets are not entirely lost. This is because each of the $Y_{rR}(k)$ frequency components transmitted is derived through a window of samples of length much larger than the packet size. Therefore the energy of one packet of speech is spread across a few packets of $Y_{rR}(k)$'s (k = 0, 1,...., N-1). The loss of one packet of $Y_{rR}(k)$'s in this case results in the loss of only part and not the whole of the energy of one packet of speech, though it also affects the following limited number of packets of speech.

Fast Algorithm for the Implementation of the Filter-Bank Analysis and Synthesis:

The implementation complexity is exorbitant if straight forward filter bank analysis and synthesis is carried out, especially when the order of the filter is high. Fortunately, a method is available [3, 4] by which the filter bank analysis process can be reduced to a very simple weighted overlap-add procedure followed by discrete Fourier transform (DFT). For our application, the DFT at the analysis and the corresponding inverse DFT at the synthesis filter banks are unnecessary. The procedure at the transmitter thus reduces to one simple processing equation given as follows:

$$U_n(q) = \sum_{r=-L}^{L-1} x(n+rN-q)\, h_o(-rN-q) \qquad (2)$$

where $q = 0, 1, 2, \ldots, N-1$ and $2LN$ is the length of the filter whose impulse response $h_o(n)$ is the sinc window:

$$h_o(n) = \sin(n\pi/N)/(n\pi/N) \qquad (3)$$

for $-LN-1 \leq n \leq LN$.

For the implementation of filter-bank synthesis, the inverse DFT can be avoided if DFT is not carried out at the analysis end. The original speech can be recovered by processing the received samples of $\hat{U}_n(q)$ by the following simple equation [3, 4]:

$$\hat{x}(n) = \sum_{m=n-LN}^{n+LN-1} \hat{U}_m((n))_N h_o(n=m) \qquad (4)$$

where "$(( ))_N$" means modulo N.

SIMULATION RESULTS AND DISCUSSION

To test the effectiveness of the proposed method for packet speech transmission, computer simulations were carried out for packet size N = 64 and two different values of 3 and 4 for L. The filter length is therefore equal to (2LN) 384 and 512 respectively. Missing packets are handled by either one of the following two different methods:

Zero substitutions : missing packets are substituted by zeros. This is the simplest of all.

Replications : when a packet is missing, the receiver simply repeats the previous correctly received packet as the present one for the recovery equation. Should two missing packets occur consecutively, the receiver will repeat the last correctly received packet twice. The same procedure is applied to any number of consecutively missing packets.

The test sentence consists of a 4-second male speech. The packet missing rate of 10 percent was simulated. The missing packets were uniformly distributed in time and happened only during speech segments which were selected through the use of a uniform random number generator. No attempt was made to simulate error burst conditions which are likely to happen in a packet data network. Our conclusions about the proposed system may therefore not be applicable under such conditions.

The performance of the system with different values of L and with the use of zero substitutions or frequency-domain replications was assessed in terms of recovered waveform comparisions (Figures 4 and 5), three-dimensional spectral amplitude plots (Figure 6) and informal subjective listening tests. Figures 4.a to 4.d and 5.a to 5.d compare the recovered waveforms of the system using different parameters and methods of handling missing packets. Compared to the conventional case of PCM packetization with zero substitutions, both Figures 4.c and 5.c show that the three missing packets are not entirely lost even with zero substitutions. Instead of the abrupt and complete loss of speech energy for a packet duration in the former case, there is residual energy in the corresponding gaps and the transition is somewhat gradual and hence improved perceptual quality of the recovered speech is expected. This is indeed confirmed by informal subjective listening tests. The improvements were significant and beyond doubt.

When the replication method was applied for missing packets, the recovered waveforms in Figures 4.d and 5.d resemble that of the original speech. As speech exhibits short-term stationarity, replications of missing frequency components will 'imitate' the short-term frequency content of the missing packets to a large extent. The recovered speech was found perceptually to have less distortions compared with zero substitutions. Compared with the original speech without missing packets, the subjective quality of the recovered speech using the replication method was only marginally inferior.

Based on informal subjective listening tests, the performance of the system using different values of L and different ways of handling missing packets, in order of merit, is as follows:

(1) L = 6 & replications
(2) L = 8 & replications
    L = 8 & zero substitutions
(3) L = 6 & zero substitutions
(4) PCM packetization with zero substitions.

When the filter length was increased to 512 (ie L = 8), the perceptual quality of the recovered speech was found to be marginally inferior to that of the system using L = 6. This is due to that fact that a missing packet now affects the following 7 packets as well and slightly more distortion in the recovered speech is expected.

CONCLUSION

A novel technique of employing filter-bank analysis and synthesis for the processing of speech before packetization for transmission

in a packet data network environment has been examined in this paper. Computer simulation results indicate that the perceptual quality of the recovered speech can be enhanced with such a system using frequency-domain replications or even with zero substitutions. Other methods of estimating the missing components to achieve better results will be studied in the future.

REFERENCES

[1] Goodman D. J., Lockhart G. B., Wasem O. J. and Wong W. C., "Waveform substitution techniques for recovering missing speech segments in packet voice communications", IEEE Trans. on ASSP, vol. ASSP-34, no.6, pp.1440-1448, Dec 1986.

[2] Wasem D. J., Goodman D. J., Dvorak C. A. and Page H. G., "The effect of waveform substitution on the quality of PCM packet communications", IEEE Trans. on ASSP, vol. ASSP-36, pp.342-348, Mar 1988.

[3] Portnoff M. R., "Implementation of the digital phase vocoder using the fast Fourier transform", IEEE Trans. on ASSP, vol. ASSP-24, no 3, pp.243-248, Jun 1976.

[4] Lee L. et al, "A new frequency domain speech scrambling system which does not require frame synchronization", IEEE Trans. on Commun., vol. COM-32, no. 4, pp.444-456, Apr 1984.

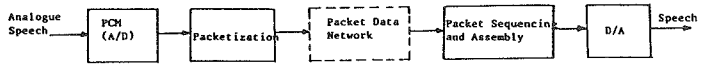

Figure 1 Packetization of PCM Speech for Packet Communications

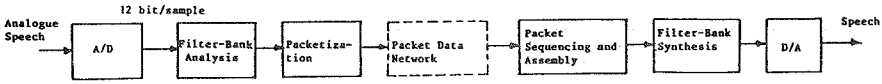

Figure 2 Block Diagram of the Proposed System for Packet Speech Transmission

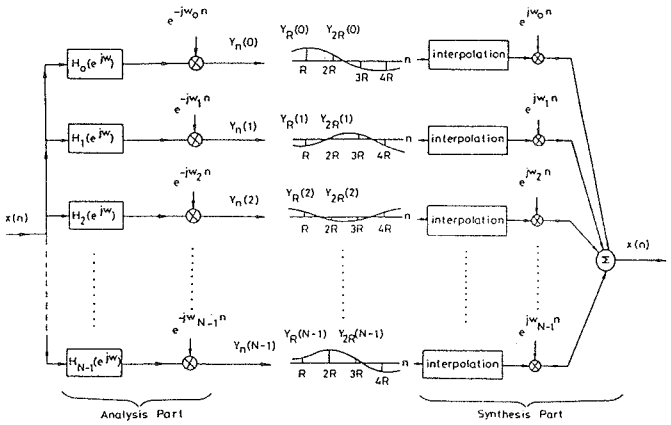

Fig 3 The analysis and synthesis methods for the ideal filter bank.

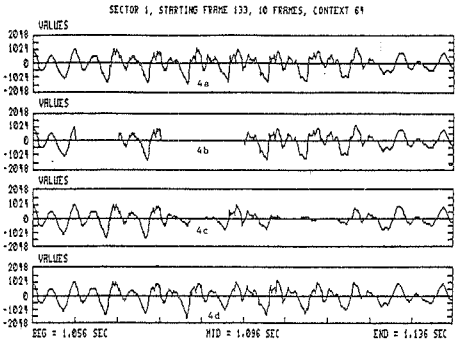

Figure 4a : Original speech waveform
4b : PCM packetization with three 64-sample packets missing
4c : Filter-bank method (L=6 zero substitutions) with three 64-sample packets missing
4d : Filter-bank method (L=6 replications) with three 64-sample packets missing
Figure 5a-5d : as Figure 4a-4d except that L=8

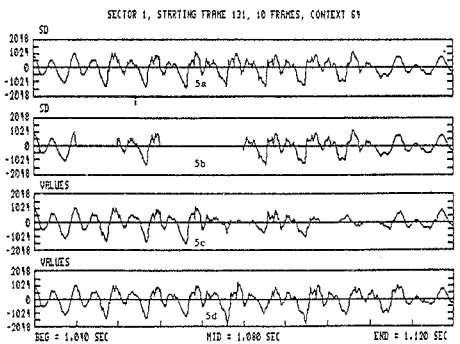

Figure 6a-6d : 3-dimensional spectral amplitude plots of Figure 4a-4d (10 frames of 64 samples each)