A NEW ERROR CONCEALMENT TECHNIQUE FOR AUDIO TRANSMISSION WITH PACKET LOSS

Alexander Stenger, Khaled Ben Younes, Richard Reng[*], *Bernd Girod*
Telecommunications Institute
University of Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany
e-mail: {stenger, younes, girod}@nt.e-technik.uni-erlangen.de
reng@vs-ulm.dasa.de

ABSTRACT

We present a new error concealment technique for audio transmission over packet networks with high packet loss rate. Unlike other techniques it modifies the time-scale of correctly received packets instead of repeating them. This is done by a time-domain algorithm, WSOLA, whose parameters are redefined so that short audio segments like lost packets can be extended. Particular attention is paid to the additional delay introduced by the new technique. For subjective hearing tests, single and double packet loss is simulated at high packet loss rates, and the new technique is compared to previous proposals by category judgment and component judgment of sound quality. Mean Opinion Score (MOS) curves show that sound distortions due to packet repetition can be reduced.

1 INTRODUCTION

In a packet network which is not designed for real-time applications, such as the Internet, audio-packets may be lost due to either congestion or excessive delay. As single or double packet loss can be assumed [1], sound quality can be improved by error concealment, which is transmitter independent and suitable for multicast. As the proposed technique exploits stationarity of speech signals, the packet length must be chosen shorter than a phoneme. We have chosen a length of 20 ms, which leads to 160 bytes per packet with 8 kHz sampling rate and 8 bit/sample. Similar values are used in [2] and [3].

There exist a number of proposals for concealment of lost speech packets [2], [3], [4]. They substitute the missing signal segment by *repeating a prior segment*, which leads to echoing or tinny sounds. To avoid these distortions, with our method the time-scale of a preceding signal segment is modified such that missing packets are covered by the *extended version of the preceding segment* (Figure 1). This operation must be performed in real-time and should preserve natural sound. An appropriate time-scale modification algorithm, WSOLA (Waveform Similarity Overlap Add), is outlined in section 2.

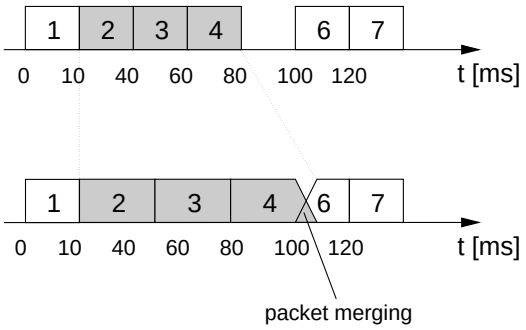

Figure 1: Error concealment using time-scale modification.

An important demand of voice communication is low delay which can only be achieved by modifying a small number of packets during the time-scaling. As shown below, the WSOLA algorithm is not designed to process such short audio segments. Therefore we will derive a modified version of WSOLA in section 3, after working out the demands on the time-scale modification algorithm. These are based on a discussion of discontinuities at the boundary between the substitute and received packets, and on the delay of the system. Experimental results are given in section 4.

2 TIME-SCALE MODIFICATION WITH THE WSOLA ALGORITHM

Successful concealment of lost audio packets implies that the time-scale modification must preserve the pitch frequency of speech (or music) signals and have low computational complexity. Such a time-scale modification algorithm with excellent sound quality already exists for continuous speech sequences, the Waveform Similarity Overlap Add (WSOLA) algorithm [5]. Figure 2 illustrates its blockwise operation: overlapping segments S_k are

[*] now with Daimler-Benz Aerospace AG, Ulm

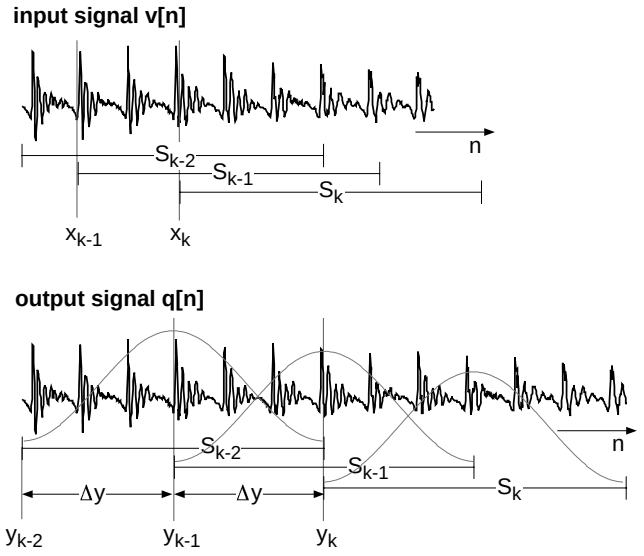

Figure 2: Time-scale modification with overlap-add techniques

extracted from the input at time instances x_k and are superimposed with less overlap in the output at positions y_k. Every part of the output arises from adding two half-segments of length Δy. To achieve smooth transitions, each segment is weighted with a hanning window $w[n]$. This procedure is given by the synthesis equation

$$q[n] = \sum_k w[n - y_k] \cdot v[n - y_k + x_k], \qquad (1)$$

where k is the step index and $w[n] \equiv 0 \; \forall \; n \notin [0; 2\Delta y - 1]$. As the output positions y_k are fixed, the window has not to be recalculated in every step.

To avoid distortion of voiced speech, the typical "peaks" should fall together when two segments are added. Therefore the segments must be cut from the input at appropriate positions x_k, which are found in a *search region* of length of at least one pitch period. As a search criterion [5] proposes maximum cross correlation between the half-segments to be added. For step k this procedure is shown in Figure 3. The boldface printed half of S_{k-1} will be superimposed with the dashed half of S_k, see output signal in Figure 2 as well. For every instance i within the search region a correlation coefficient

$$c_i = \sum_{j=0}^{\Delta y - 1} v[i + j] \cdot v[x_{k-1} + \Delta y + j], \qquad (2)$$

has to be calculated.

The positions of the search regions in the input signal have great influence on the capacity of the algorithm. As the WSOLA-algorithm was developed to provide a certain extension factor α using long speech sequences, the starting positions x_{k-1} and x_k of subsequent segments have to be spaced on average by $\Delta x = \frac{\Delta y}{\alpha}$. This is achieved with search regions centered around $x_{k-1} + \Delta x$, as shown in Figure 4.

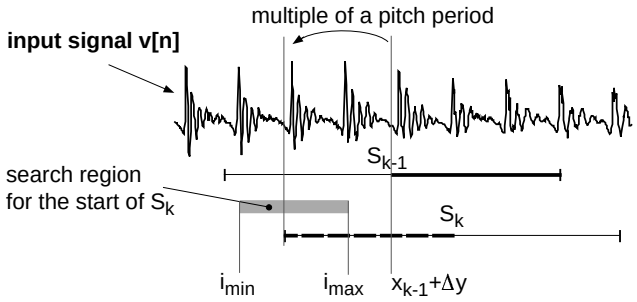

Figure 3: Finding a "synchronized" overlap segment S_k.

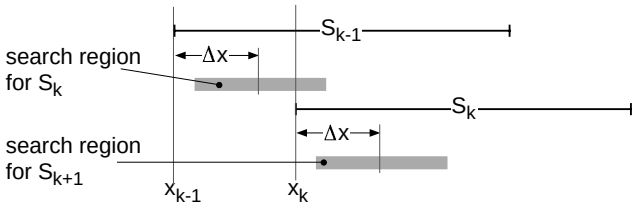

Figure 4: Positions of the search regions.

Thus every search region is placed relative to the position found for the previous segment. As the positions of the search regions depend on the results of previous search operations, the number of required input samples may vary over a wide range, if a small number of segments is used. This is a disadvantage for our application, where the number of input samples is limited. Therefore we have to modify the positioning of the search regions.

3 APPLICATION OF WSOLA FOR ERROR CONCEALMENT

Using time-scale extension for error concealment leads to problems that will be addressed in the first subsection. Then, in 3.2, the demands on time-scaling algorithms are summarized and in subsection 3.3 a new parameter selection scheme for the WSOLA algorithm is derived.

3.1 Problems

To avoid discontinuities at the boundary between the substitute and next correct packet, a merging technique as proposed in [2] is used with an overlap of 10 samples (see Figure 1). Thus, clicking sounds can be suppressed if the fundamental frequency of voiced speech is constant during all modified packets including the gap or if the speech segment is unvoiced. Otherwise, a phase difference occurs that cannot be concealed by packet merging [3]. To overcome this, the idea of "phase matching"

as proposed in [3] could be applied to our scheme. This is currently under investigation and will not be considered here.

Another problem is the additional delay caused by error concealment. We assume that the error concealment routine is started when the size of the gap, d_{lost}, is known, i.e. the first correct packet after the gap has just arrived, and that the time d_{calc} passes before the first audio output is ready. Techniques like [2] - [4], that do not modify packets preceding the gap, introduce an extra delay of $d_{calc} + d_{lost}$. Our technique leads to a delay as long as the duration of the audio data modified by the algorithm, d_{mod}, in addition to $d_{calc} + d_{lost}$.

3.2 Demands

The conditions for minimum delay depend on the way the data is sent to the output. With common workstations, blocks of audio samples can be written to the output queue at any time and are played out continuously. This means that d_{calc} should be as small as possible. This is achieved by keeping the number of segments (i.e. correlation calculations) small. The second parameter to minimize is d_{mod}. If we assume that the audio output blocksize is 1 packet, modifying should begin at a packet boundary, and the number of modified packets should be small. A typical example is shown in Figure 5.

Therefore packet error concealment makes different demands on a time-scale modification algorithm than the ones described in section 2: not an average extension factor has to be achieved, but a given short speech segment shall be extended to a guaranteed minimum length using only a few segments. To preserve the excellent audio quality of the WSOLA algorithm, multiple repetition of a part of the input sequence and repetition of a segment taken too far from the left should be avoided, as the first leads to tinny sounds and the latter may cause echoes. The parameter selection proposed in the following leads to a tradeoff between sound quality and delay.

3.3 New parameter selection scheme

Figure 5 shows how the loss of packet 4 is concealed using three correctly received ones. Only packets 2 and 3 containing l_{in} samples are modified, so the delay is 3 packets plus d_{calc}. With a packet size of 160 samples and 10 extra samples for packet merging, the algorithm must produce at least $l_{out} = 490$ samples. To obtain a given minimum output length, first the number of segments N is determined and then the segment length $L = 2\Delta y$ is calculated such that it satisfies

$$\frac{L}{2}(N+1) > l_{out}. \tag{3}$$

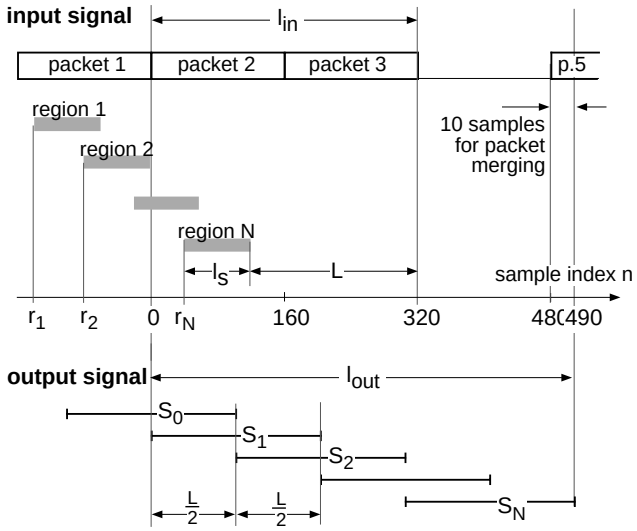

Figure 5: Modified WSOLA algorithm.

To allow the extraction of a full segment in the last step (N), the end of search region N is placed L samples left from the end of the input signal. Thus the resulting output length is between $(N+1)\frac{L}{2}$ and $(N+1)\frac{L}{2} + l_s$, where l_s denotes the length of the search regions.

The positioning of the other search regions is a delicate problem, if natural sound shall be preserved. Increasing overlap may cause multiple repetition of the same segment, which leads to tinny sounds. If the first search region is placed too far to the left, a part of packet 1 may be entirely repeated resulting in echo. With equally spaced search regions, the choice of region 1 allows a tradeoff between echoing or tinny sound distortions. As the search regions overlap more with higher extension factors $\frac{l_{out}}{l_{in}}$, the first search region has to be placed more to the left in this case to keep both kinds of distortion balanced. An experimentally found formula for the start of search region 1 is

$$r_1 = -l_s - 80(\frac{l_{out}}{l_{in}}). \tag{4}$$

As a result the sound quality for extension factors up to 2 is acceptable, if a sampling rate of 8 kHz and a segment length of 200-300 samples is used. A segment length in this range is achieved if N is chosen as

$$N = \lfloor \frac{l_{out}}{100} - 1 \rfloor \tag{5}$$

and eqn. (3) is satisfied setting

$$L = 2 \cdot \lceil \frac{l_{out}}{N+1} \rceil. \tag{6}$$

Calculating N with eqn.(5) leads to a relatively small number of steps and therefore to low computational

complexity. A smaller number of steps should be avoided as this leads to larger segments which might produce echoing sounds.

4 EXPERIMENTAL RESULTS

The new Time-scale Modification technique (TM) was compared to Silence Substitution (S), Pattern Repetition (PR), Pitch Waveform Replication (PWR) by subjective hearing tests. We simulated single packet loss by suppressing one packet within five and double packet loss by suppressing two packets within seven. Thirteen non-expert listeners were asked to judge overall quality as well as the presence of the distortion components "tinny, metal", "interrupted, clicking" and "echoing, reverberating". We used 48 test conditions (male and female speakers) sampled at 8 kHz. The "worst case" anchor contains clearly audible distortions of all the three types. The Mean Opinion Score (MOS) of the overall

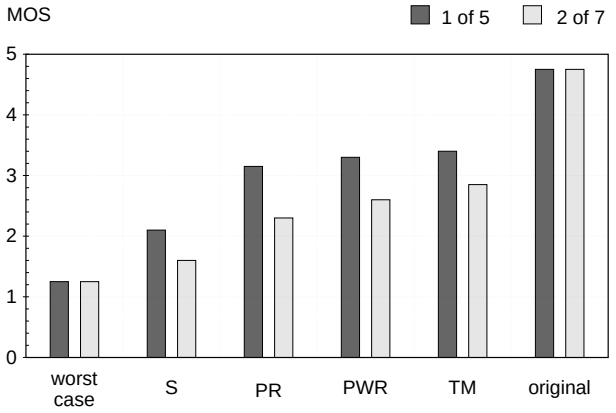

Figure 6: Mean Opinion Scores of overall quality.

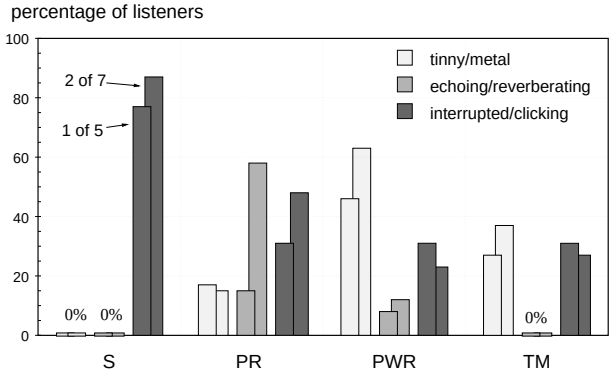

Figure 7: Component Test.

quality (Figure 6) shows a quality enhancement of TM compared to all other techniques.

The values in Figure 7 indicate how many test persons noticed a specific distortion component. It can be seen that the echoing sound produced by PR is eliminated completely and the tinny sound of PWR is reduced by the new TM technique. The component "interrupted/clicking" is still noticeable as with the previous techniques, but could be reduced considerably by phase matching.

5 CONCLUSION

A new error concealment technique for lost audio packets based on time-scale modification has been proposed. First the WSOLA time-scaling algorithm was explained and it was shown that it cannot extend short audio segments to a given length. To overcome this problem, a new parameter selection scheme was derived that leads to low delay and low computational complexity. Experiments have shown that typical disturbance components of other techniques are reduced and overall quality is improved.

References

[1] Jean-Chrysostome Bolot, Hugues Crepin, Andres Vega Garcia, "Analysis of Audio Packet Loss in the Internet," *Proc. of 5th Int. Workshop on Network and Operating System Support for Digital Audio and Video*, pp. 163-174, Durham, April 1995.

[2] Ondria. J. Wasem, David. J. Goodman, Charles A. Dvorak, Howard G. Page "The Effect of Waveform Substitution on the Quality of PCM Packet Communications," *Trans. on ASSP*, vol. 36, No. 3, pp. 342-347, March 1988.

[3] R. A. Valenzuela, C. N. Animalu, "A New Voice Packet Reconstruction Technique," *Proc. ICASSP 89*, pp. 1334-1336.

[4] David. J. Goodman, Gordon B. Lockhart, Ondria. J. Wasem, Wai-Choong Wong, "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications," *IEEE Trans. on ASSP*, vol. 34, No. 3, pp. 342-347, Dec. 1986.

[5] Werner Verhelst, Marc Roelands, "An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech," *Proc. ICASSP 93-II*, pp. 554-557, April 1993.