PITCH RESYNCHRONIZATION WHILE RECOVERING FROM A LATE FRAME IN A PREDICTIVE SPEECH DECODER

Kyle D. Anderson, Philippe Gournay

VoiceAge Corporation
750, Chemin Lucerne
Montréal (Québec) Canada H3R 2H6

Philippe.Gournay@USherbrooke.ca

10.21437/Interspeech.2006-68

ABSTRACT

The concealment procedure used by CELP speech decoders to regenerate lost frames introduces an error that propagates into the following frames. Within the context of voice transmission over packet networks, some packets arrive too late to be decoded and must also be concealed. Once they arrive however, those packets can be used to update the internal state of the decoder, which stops error propagation. Yet, care must be taken to ensure a smooth transition between the concealed frame and the following "updated" frame computed with properly updated internal states. During voiced or quasi-periodic segments, the pitch phase error that is generally introduced by the concealment procedure makes it difficult and detrimental to quality to use the traditional fade-in, fade-out approach. This paper presents a method to handle that pitch phase error. Specifically, the transition is done in such a way that the natural pitch periodicity of the speech signal is not broken.
Index Terms: speech coding, robustness, late packets

1. INTRODUCTION

In today's world, with quicker and more widely available Internet connections, voice over IP communication has become a reality. The same speech codecs used in circuit-switch networks now face new challenges related to packet switched communications.

When transmitting speech over packet networks, a packet consists of one or several speech frames. Lost packets and packets that arrive after their playout time cannot be decoded. A concealment procedure is called instead of the standard decoding procedure in order to replace the missing speech or audio segment. Unfortunately, this concealment procedure is not perfect and an error is introduced in the concealed segment. Moreover, in predictive codecs, this concealment procedure does not correctly update the internal state of the decoder. Thus the error introduced by the concealment generally propagates in the following segments.

This paper provides a solution to recovering from packets received late in the context of a predictive codec. More specifically the focus is on an algorithm to recover from the late reception of a voiced segment of speech in a CELP-based codec without creating unpleasant artifacts.

1.1. Prior solutions

The idea to use late packets in order to reduce concealment induced error propagation when transmitting voice over packet networks has been around since at least 1983 [1]. When a packet is not lost but simply delayed, its contents can be used to update "a posteriori" the internal state of the decoder. This limits, and in some cases stops, the error propagation caused by concealment [2]. Fig. 1 illustrates the steps taken to update the internal state of the decoder when a late packet containing one frame is received.

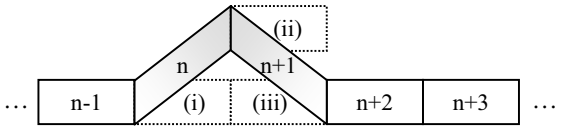

Fig. 1: Chronology of performing an update

Frame n does not arrive before its playout time and is thus concealed creating an error in the internal state of the decoder. When frame n arrives before the playout time of frame n+1 it is decoded, segment (i), as if it had arrived on time. Frame n+1, which is assumed to have arrived on time, is then decoded twice. First as if frame n had not arrived, segment (ii), and then as if frame n were on time, segment (iii). Segment (ii) and (iii) are now combined to form the output speech of frame n+1 and eliminate further error propagation due to frame n arriving late. While the error propagation created by concealment only lasts two frames (here, frames n and n+1), great care must be taken to ensure a smooth transition back to the correct signal during frame n+1. It is in this frame that recovery from the late frame takes place.

The method presented in [2] for recovering after a late frame uses simply a fade-in, fade-out window on segments (iii) and (ii), respectively. This technique was applied in the excitation domain. While this smoothing technique works in some cases, it does not take into account all types of errors introduced by concealment.

1.2. Present problem

The concealment procedure of predictive speech decoders generally introduces a pitch phase error during voiced segments. This problem, though largely overlooked until recently, has serious implications for the recovery of the decoder [3]. It also makes it difficult and detrimental to quality to use the traditional fade-in, fade-out approach when passing from the concealed output segment (segment (ii) in Fig. 1) to the following "updated" output segment computed with a properly updated internal state (segment (iii) in Fig. 1).

Fig. 2 shows, in the excitation domain, what happens when that approach fails. Signal 1 is the correctly decoded excitation signal with no errors. Note that the third frame of signal 1 corresponds to segment (iii) in Fig. 1. Signal 2 is the excitation when the second frame is considered as lost and concealed. Here the third frame corresponds to (ii). Notice that the third frame of signal 2 has fallen out of phase with signal 1. Signal 3 is the excitation signal when the second frame is received late and the fade-in, fade-out mixing technique is applied to the third frame. When these two signals are mixed, two pitch pulses occur side-by-side in the resulting signal (indicated by an arrow in Fig. 2). As a result, the natural pitch periodicity is lost and an unpleasant break occurs in the energy of the final, synthesized signal.

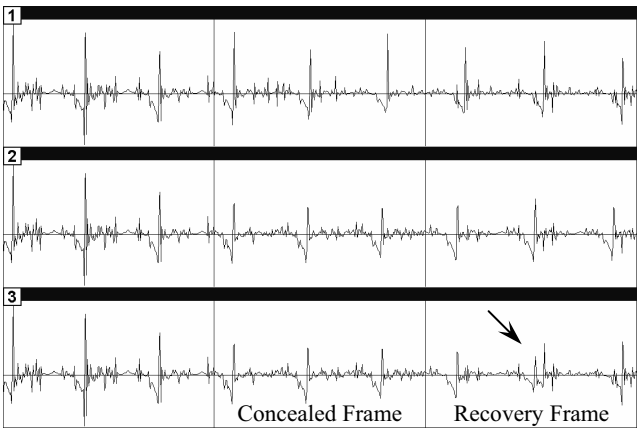

Fig. 2: Signal 1, no error; Signal 2, 2nd frame lost; Signal 3, 2nd frame late, recovery method from [2]. The arrow points to the effect of the pitch phase error.

1.3. Solution

To avoid a pitch phase distortion after the reception of a late voiced frame, the concealed and updated segments (ii) and (iii) in Fig. 1 must be realigned before mixing. Since the pitch may vary between pulses, it is only possible to align one pulse at a time; however, this is sufficient. When the two segments are properly aligned, it becomes possible to switch, at a point with low energy, from the concealed segment to the updated segment without creating a discontinuity in the pitch contour of the speech signal.

The remainder of this paper will focus on the recovery effort in the frame following the reception of a late voiced frame. In relation to Fig. 1, the solution will detail how to combine segments (ii) and (iii) such that unwanted artifacts are minimized during this recovery process.

2. PROPOSED RESYNCHRONIZATION METHOD

The terminology used in the algorithm is briefly explained in the following subsection 2.1. Late unvoiced frames are treated as described in subsection 2.2. The algorithm used for late voiced frames is described in subsection 2.3. Finally, alternative realizations are presented in 2.4. The presentation is in the context of the VMR-WB decoder but could be applied to any CELP-based decoder. Note the recovery algorithm does not affect the encoder.

2.1. Terminology

In the following, the term "good" excitation will refer to segment (iii) in Fig. 1 above and "bad" excitation to segment (ii). The symbol "T_0" will be used to represent the pitch period; it will always refer to the pitch of the first subframe in the good excitation unless otherwise noted. In VMR-WB, T_0 is a known parameter transmitted in the coded speech packet. For other codecs, some sort of pitch analysis may be necessary to its value at the decoder.

2.2. Late unvoiced frames

Unvoiced speech is characterized by a lack of periodicity. This is a result of sound produced without vibration of the vocal cords. This lack of periodicity means that when an unvoiced frame is concealed, pitch desynchronization does not pose a problem. However, an error is still introduced into the signal. Since the human ear is not sensitive to phase differences in unvoiced speech due to the noise-like nature of the signal, the fade-in, fade-out window is very effective.

2.3. Late voiced frames

As mentioned in the introduction, when concealment is performed on voiced or quasi-periodic speech segments, a pitch phase difference is generally introduced. This means that when a voiced frame is received late, the recovery should not be effectuated in the same fashion as for unvoiced frames. When this is done, irregularly-sized pitch periods, which are detrimental to speech quality, are introduced. Also, as seen in the introduction (Fig. 2), simply using a fade-in, fade-out window exposes the risk of having

two pitch pulses side-by-side or even canceling out each other. The solution proposed here solves these problems by aligning the good and bad excitations so as to avoid any irregular pitch periods or energy levels.

Three main steps are necessary to solve the problem of resynchronizing two out-of-phase voiced excitation signals. First, a glottal pulse to be used in the resynchronization must be found (Fig. 3 block 1); this can be done in either the good or bad excitation. Second, the offset which maximizes the correlation between the good and bad excitation is found (Fig. 3 block 2). And finally, a minimum energy point must be found; this is where the switch from the bad to good excitation is made (Fig. 3 block 3). If the energy level of the pulse or the strength of the correlation is not within acceptable limits, as described below, the overlap-and-add from [2] is used to combine the excitations (Fig. 3 block 5).

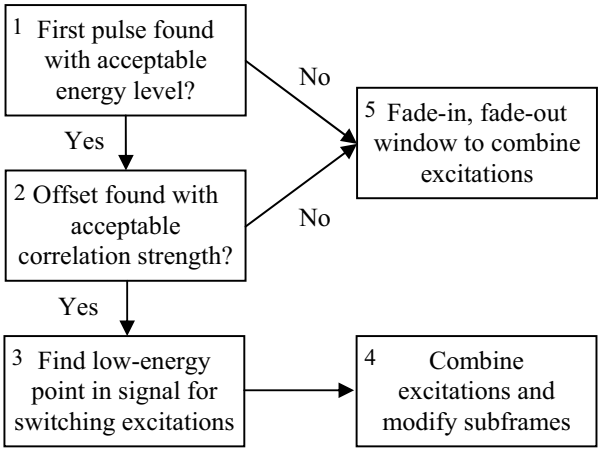

Fig. 3: Recovery steps for voiced frames

To find the pulse, a small window (e.g. 5 to 10 samples) is used to scan over a section of the excitation slightly larger than T_0 (illustrated in Fig. 4). Searching over a section slightly larger than T_0 protects against finding half-pulses lying on the extremities of the pitch period. The point where the window demonstrates the highest energy (*P* in Fig. 4) should contain the pulse. To verify that the pulse found is an actual pulse, a verification of the energy it contains is performed. This can be done by comparing the energy contained in the pulse (the maximum energy found with the scanning window) to the energy in a fixed number (that should be no greater than the minimum pitch length to avoid including another pulse) of samples surrounding the pulse. A lower threshold for this energy ratio is to be found experimentally.

The glottal pulse found to work best is the first pulse in the good excitation. The first pulse in the bad excitation was also tried, but the concealed pulses are often less distinct making them more difficult to isolate. For the rest of the discussion, it will be assumed that the good excitation was used to find the initial glottal pulse.

This pulse is dragged across the bad excitation, using another small window to calculate the cross-correlation between the good and bad excitations. The amount of shift (in samples) that maximizes the correlation, gives the offset between the two signals (Fig. 3 block 2). An appropriate lower threshold, found experimentally, is set on the correlation to ensure consistent and accurate results.

Cross-correlation is usually normalized by the energy contained in both signals, but it was found that finding the maximum when only normalizing by the energy in the good excitation worked best in matching the pitch pulses and retaining their natural periodicity. Equations (1) and (2) below show the offset calculation used in this specific implementation. Notice that the cross-correlation calculation as in equation (1) may be greater than one.

$$\hat{j} = \arg\max_{j} \left(\frac{\sum_{i=0}^{i=W-1} good[P+i]\,bad[P+i+j]}{\sum_{i=0}^{i=W-1} good[P+i]^2} \right) \quad (1)$$

and

$$0 \le j < T_0 \text{ and } j \le FL - P - W \quad (2)$$

Here, *P* represents the index of the beginning of the window containing the glottal pulse found in the good excitation; *W* represents the size of the correlation window used; *j* represents the amount the pulse is shifted across the bad excitation; and $\hat{j}$ represents the offset between the two signals. Note that the bounds on *j* may be changed to search both forwards and backwards in the bad excitation if desired. A graphical interpretation of these parameters can be found in Fig. 4.

Once the offset is found, the existence of a pulse in the bad excitation may be verified, as was done when the first pulse was found in the good excitation.

The final step in the recovery process is to combine the good and bad excitations. A point where the energy is low in the two signals to be combined should be found. Since they now have one pulse aligned, the minimum energy point (*m* in Fig. 4) should be found either just before or just after this pulse. Also, it need only be found in one of the two signals. This point correspond to the position where the change from the bad to good signal takes place. Note in Fig. 4 that the end of signal 3 corresponds exactly to the part of signal 1 that is at the right of point *m*.

When the offset found by correlation is non-zero, the recovery frame will be either longer or shorter than a standard-sized frame. In the VMR-WB, post-filtering on the signal is performed on a subframe-by-subframe basis; thus, the sum of the subframe lengths must correspond to the length of the entire frame. At the very end of the recovery process, new subframe lengths must be calculated to satisfy this constraint.

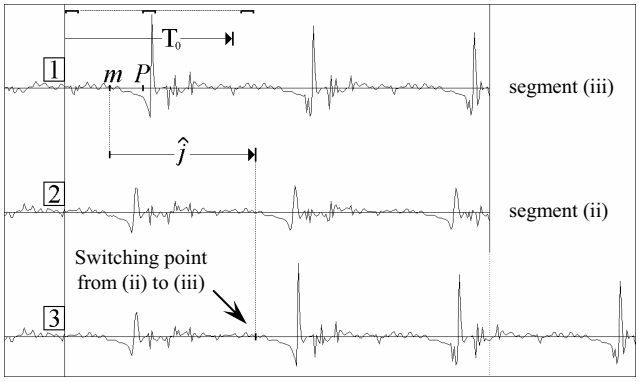

Fig. 4: Signals and parameters used for resynchronization

2.4. Alternative realizations

Section 2.3 described one way to determine the pitch phase error. This error may also be determined using PSOLA (Pitch-Synchronous Overlap and Add) by first finding the pitch marks in (ii) and (iii), then comparing the positions of those marks. Alternatively, the error may be found by first determining the position of the last pitch mark in frame n-1, then using the concealed and actual pitch values of frame n to deduce the pitch marks positions in (ii) and (iii).

In section 2.3, the pitch phase error was compensated by delaying (iii). As soon as one pulse in (ii) is "in phase" with another in (iii), it is possible to switch rapidly from one segment to the other without breaking the periodicity. Because a delay is applied to (iii) however, the resulting "recovery" frame is longer than usual. In some applications this is acceptable and is even desirable (i.e. with adaptive jitter buffering, a longer frame increases the playout delay which in turn reduces the probability of receiving another late frame). In the other applications were a constant frame duration is required, a "recovery" frame of normal length may be obtained by progressively shifting back to their normal position the pulses in the "recover" excitation.

3. RESULTS

We implemented this new resynchronization method in the VMR-WB [4] decoder. The complexity and memory requirements are only slightly increased with respect to the original method presented in [2]. Objective and subjective results are presented below.

3.1. Illustration of the proposed method

Fig. 5 is a screen shot demonstrating the efficiency of the proposed method. Signal 1 is the decoded speech signal with no errors. In signal 2, the third frame is received late and is processed according to [2]. In signal 3, the third frame is received late and processed according to the proposed resynchronization method. In the fourth frame of signal 2, the recovery effort clearly creates a break in the energy of the signal, whereas with the improved method used in signal 3 the energy continuity is maintained.

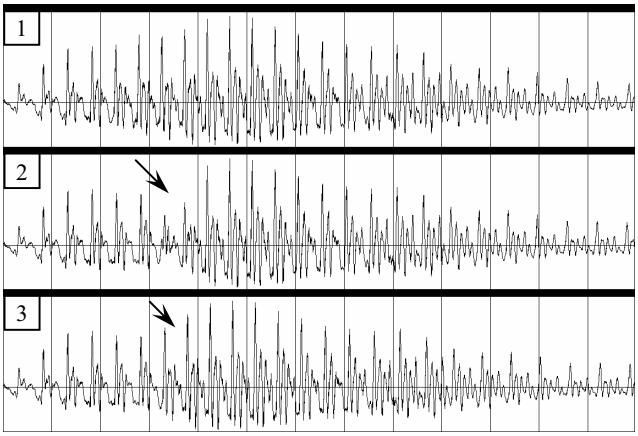

Fig. 5: Signal 1, no errors; Signal 2, 3rd frame late, update according to [2]; Signal 3, 3rd frame late, update using presented method. The arrow indicates the recovery frame.

3.2. Subjective improvement

The conditions necessary for this method to be used do not occur very often. However, when cases are presented, the method described corrects audible artifacts during the recovery effort. Specifically, highly localized but strongly annoying clacks caused by either energy losses or breaks in the pitch contour are avoided. We did not perform any formal listening test, but audio samples that demonstrate the improvement are available upon request.

4. CONCLUSION

During voiced segments, the concealment procedure used by predictive speech decoders often produces a pitch phase error. When a late packet is used to update the internal state of the decoder, this error makes it difficult to use the classical fade-in, fade-out technique to smooth the transition between the concealed output segment and the following "updated" output segment. This paper presented a method to solve this problem by resynchronizing the excitation signals corresponding to the concealed frame and to the following frame where error recovery is performed. The transition is therefore done in such a way that the natural pitch periodicity of the speech or audio signal is not broken.

5. REFERENCES

[1] W. A. Montgomery "Techniques for Packet Voice Synchronization", IEEE Journal on Selected Areas in Communications, Vol SAC-1, No. 6, December 1983.

[2] P. Gournay, F. Rousseau, and R. Lefebvre, "Improved packet loss recovery using late frames for prediction-based speech coders", ICASSP'2003, Hong Kong, April 6-10, 2003.

[3] M. Chibani, R. Lefebvre, P. Gournay, "Resynchronization of the Adaptive Codebook in a Constrained CELP Codec after a frame erasure", ICASSP'2006, Toulouse, March 14-19, 2006.

[4] M. Jelinek, et al., "On the architecture of the CDMA2000 variable-rate multimode wideband (VMR-WB) speech coding standard", ICASSP'2004, Montréal, May 17-21, 2004.