SIGNAL MODIFICATION FOR CODING PURELY VOICED SECTIONS IN A WIDEBAND ACELP SPEECH CODER

Mikko Tammi[1] and Milan Jelinek[2]

[1]Tampere University of Technology, Digital and Computer Systems Laboratory, Tampere, Finland
[2]University of Sherbrooke, Department of Electrical Engineering, Sherbrooke, Quebec, Canada

ABSTRACT

In narrowband CELP coding, signal modification is often employed to improve pitch prediction at lowest bit rates. In this paper, we extend signal modification for wideband speech coding. A method for identifying, modifying, and coding purely voiced speech frames in a variable bit rate wideband ACELP coder is presented. Signal modification enables coding of purely voiced frames with a low bit rate mode. Modification is performed frame and pitch synchronously preserving the original time scale at the end of each frame. Since no modification is allowed at the frame end, the delay parameter of pitch prediction is prone to oscillations. These oscillations can be reduced by choosing a proper delay contour for interpolating the parameter over the modified frame We also demonstrate that the signal modification algorithm can be employed for detecting reliably frames suitable for signal modification. Informal listening tests indicate good performance for the proposed method.

1. INTRODUCTION

In the conventional code excited linear prediction (CELP) coding, pitch prediction is usually performed on a subframe basis. A delay parameter of long term prediction, which maps the past excitation signal into the present, is assigned to each subframe. For good performance, subframes should not be much longer than 5 ms. However, at the lowest bit rates transmitting the delay parameter this frequently requires a significant proportion of the available bit budget. The performance of the low bit rate CELP coders can be improved by allowing small modifications to the input signal [1–3]. Signal modification adapts the evolution of the pitch cycles in speech signal to fit the coding model, enabling to transmit only one delay parameter per frame. CELP coders utilizing this principle are often referred to as generalized analysis-by-synthesis coders.

Previously signal modification has been utilized in narrowband speech coding, such as in the RCELP concept [2]. In this paper, signal modification is used for efficiently coding purely voiced frames at a low bit rate in a source controlled ACELP wideband coder. The coder operates at a sampling rate of 16 kHz, and comprises full and low rate modes at 12.65 and 7.6 kbps, respectively. A rate determination functionality is included in the signal modification algorithm as depicted in Fig.

The email addresses of the authors are mikko.tammi@tut.fi and milan.jelinek@courrier.usherb.ca. The work was done during M. Tammi's one year visit to the University of Sherbrooke. This work was supported by VoiceAge Corp.

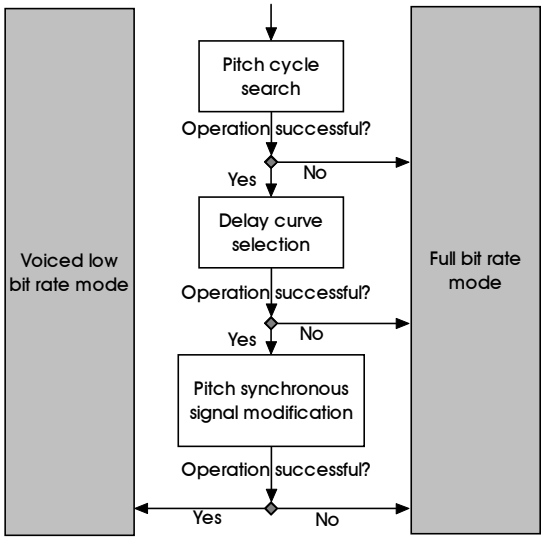

Figure 1. Simplified block diagram of the proposed signal classification and modification system.

1. The full rate mode has been adopted as such from the AMR-WB standard [4, 5] for coding the frames not classified as purely voiced. The low rate mode has been implemented for the experimental part of this paper. In sequel we concentrate on describing identification, modification, and coding of the purely voiced frames.

The proposed method does not modify the transition segments of the signal, such as voiced onsets, because the perceptual quality may easily be degraded. In purely voiced sections, pitch cycles typically change relatively slowly, and small modifications suffice to adapt the signal to the coding model. This reduces the risk of causing artifacts. Only the frames classified as purely voiced are modified, other frames are kept intact. Thus the processing is frame synchronous. If no modification is allowed in the next frame, whose class is yet unknown, the signal modification in the current frame has to be done preserving signal continuity at the frame boundary. Hence, the samples close to frame boundaries cannot be shifted, i.e. no time asynchrony is allowed for the last samples in the frame. When no time asynchrony is allowed at the end of the frame, the resulting pitch in the modified signal tends to oscillate around the true original pitch. The delay contour used for interpolating the delay value over the modified frame has to be controlled properly to avoid these oscillations.

In the proposed method, signal modification is done pitch synchronously, i.e. adapting one pitch cycle segment at a time. Segments are limited by the frame boundaries. A simplified block diagram of the proposed signal modification method is presented in Figure 1. The algorithm starts by locating individual pitch cycles. Based on the search result, the frame is divided into segments each containing one pitch cycle. Next, a delay contour is selected for interpolating the delay parameter of the long term predictor over the frame. The pitch of the modified signal will follow this contour. The delay contour has to be chosen considering time asynchrony limitations. Based on the selected delay contour a target signal can be formed for adapting pitch cycle segments. Finally, pitch cycle segments are shifted one by one to maximize the correlation with this target signal. To keep the complexity at a low level, no time warping is applied in shifting the segments.

The signal modification method as such contains also an efficient classifier for purely voiced segments. The modification algorithm is enabled in each frame not classified as unvoiced. The modification algorithm provides several indicators on the periodicity of the current frame. If an indicator exceeds its allowed limits, it can be concluded that the current frame is not suitable for modification, and thus the original signal is maintained. Therefore we discuss signal classification in this paper along with he signal modification algorithm. The organization of this paper is as follows. Section 2 presents means for dividing a speech frame into pitch cycles. The selection of the delay parameter to be transmitted and the delay contour for interpolation are discussed in Section 3. In Section 4, pitch synchronous signal modification itself is described. Some results on the performance of the proposed method are given in Section 5. Conclusions of this paper are drawn in Section 6.

2. PITCH CYCLE SEARCH

The first phase of the signal modification is to divide frame into segments each containing one pitch cycle. Usually the most apparent part of the pitch cycle is the pitch pulse. The pitch pulses of the frame are located first, and this information is used in partitioning the frame. Pitch pulses in current frame are searched from the weighted speech signal $w(t)$ composed with the filter

$$W(z) = \frac{A(z/\gamma_1)}{1-\gamma_2 z^{-1}}, \quad (1)$$

where $A(z)$ is the LP analysis filter and γ_1 and γ_2 are the weighting coefficients. Before pulses from the current frame are sought, the last pitch pulse in the residual signal $r(t)$ of the previous frame is first searched. By low-pass filtering and squaring the signal, a pitch pulse typically stands out as the maximum value. Instead of finding the very exact pitch pulse position, it is more important that the maximum energy section of the signal is situated close to the detected pitch pulse.

A pulse prototype is extracted in the previous frame using the weighted speech signal, and it is employed to search for the pitch pulses in the current frame. As the position T_p of the last pulse in the previous frame has been found in the residual signal, a segment centered to this position is extracted from the weighted speech to be used as the prototype. The length of the model is fixed to 21 samples. The detected pulse is placed to the middle of the model. The first pulse of the present frame should occur approximately at instant $T_p + p(T_p)$, where $p(t)$ is the interpolated open-loop pitch estimate at instant t. The pulse prototype is correlated with the neighborhood of the predicted pulse position to find the exact position of the next pitch pulse. The procedure is repeated until the end of the frame is achieved. The integer resolution suffices for the other pulse positions except the last one, for which a fractional resolution is used. The last pulse is searched more accurately, since the exact distance between the last pulses in the previous and the current frames is needed in determining the delay parameter to be transmitted.

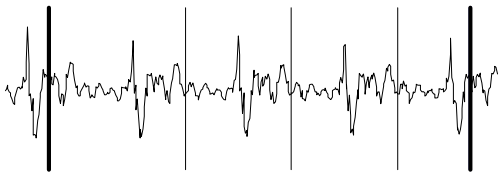

Figure 2. An example of dividing speech frame into segments. Frame boundaries are denoted with thick lines.

In signal modification, detected pitch cycles are shifted independently to adjust the signal to the target signal. It is important that segment boundaries occur at low power sections of signal which have small perceptual importance, because the actual shape of the signal is modified only on these areas. We ended up to a solution in which the segment boundaries are placed approximately to the middle between two pitch pulses. The exact position of boundary is selected by finding instant with smallest energy. In the beginning and the end of the frame segments are limited by the frame boundaries. At both ends of the frame segments are selected such that each segment contains exactly one pitch pulse. An example of dividing speech frame into segments is given in Figure 2.

As was mentioned before, signal modification algorithm can also be used for recognizing purely voiced sections of speech. Especially the above described pulse search process gives a lot of useful information about the periodicity of current frame. Certain requirements can be set for the strength of the correlation between detected pulses and the pulse model. In addition, major alternation is not allowed in the distance between detected pulses, and all those distances should also be close to the open loop pitch estimate. If any of these measurements is not within allowed limits, modification process is continued no further and the frame is coded in the full bit rate mode.

3. DELAY SELECTION

Generally the main motivation for the signal modification is that only one delay value per frame have to be quantized and transmitted to the decoder. However, special attention has to be paid to this single parameter. The delay value of the current frame defines together with previous frame's value the evolution of the pitch cycle length, as well as effects to the time asynchrony in the modified signal. In most implementations some time asynchrony is allowed at frame boundaries, as for example in RCELP coding [1], in which the value of delay parameter can be chosen straightforwardly using the open loop pitch estimate (with certain limitations). However, in our

implementation, basically no shift is allowed for the last segment in the frame. To confirm that there is no time asynchrony at the end of the frame, the delay contour must map the most important signal features at the end of the frame to corresponding features at the end of the previously synthesized frame. In practice this implies that the last pitch pulse in the frame should be linked to the last pitch pulse in the previous frame through the delay contour. In mathematical form this can be expressed as follows: let T_p and T_c be the instants of the last pulses in previous and current frames, respectively. In addition, let us assume that there are j pitch pulses between these two instants. The delay contour $d(t)$ should be selected as follows: with initialization $t_t = T_c$, repeating equation $t_t = t_t - d(t_t)$ in total $j + 1$ times, t_t should end up to instant T_p or very close to it. Only if this holds adaptive codebook can be utilized efficiently without time asynchrony at the end of the frame.

As important as the delay value at the end of the frame is the shape of the delay contour. Typically, simple linear interpolation has been employed between transmitted delay values [1]. However, in our case with no time asynchrony allowed at the end of the frame, linear interpolation tends to result in a framewise oscillating delay contour. Oscillating delay value implies oscillating pitch cycle length, which in turn rapidly degrades the quality of modified speech. The evolution and the amplitude of the oscillations are in a complicated manner related to the position of the last pitch pulse in a frame. If the last pulse is located far from the frame boundary, the frame is more prone to increase the amplitude of the oscillation. According to our experiments the amount of the oscillation can be significantly reduced simply by using a delay contour which grows linearly in the beginning of the frame and remains constant at the end:

$$d(t) = \begin{cases} (1-\alpha(t))d_{n-1} + \alpha(t)d_n & t_{n-1} < t < t_{n-1} + \sigma_n \\ d_n & t_{n-1} + \sigma_n \le t \le t_n \end{cases}, \qquad (2)$$

where t_n and t_{n-1} are the end instants of current and previous frames, d_n and d_{n-1} are the corresponding delay values, $\alpha(t) = (t - t_{n-1}) / \sigma_n$ and $t_{n-1} + \sigma_n$ is the instant after which the delay contour remains constant. The variable σ_n can either be held constant or varied as a function of d_{n-1}. Notice that depending on the class of the previous frame, d_{n-1} can be either delay value at the end of the frame (signal modification mode) or the delay value for the last subframe (full bit rate mode).

There is no simple explicit method to solve for d_n. Instead, several values have to be tested to find the best solution. However, the search is not complex and since the value of t_t (after $j + 1$ iterations) changes consistently as a function of d_n, it is quite straightforward to find accurate value for d_n. At the sampling rate of 12.8 kHz, high frequency female voices require delay values with accuracy of 1/4 sample whereas for lower female voices and for male voices resolution of 1/2 sample is enough. These requirements can be fulfilled with 9 bits for the delay quantization per frame.

To illustrate the difference between linear and proposed interpolation method an example is given. An artificial speech signal with constant pitch cycle length of exactly 52 samples was generated. The signal was divided into 20 ms frames each comprising 256 samples. The delay value in the beginning of the very first frame was intentionally selected incorrectly to 54 samples to illustrate the effect of pitch

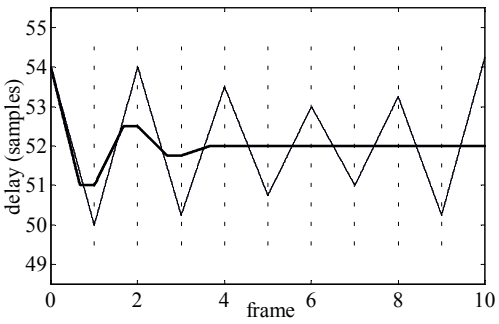

Figure 3. Example on the delay contour with the traditional linear interpolation method (thin line) and with the proposed new interpolation method (thick line).

estimation errors typical in speech coding. Then, using both linear interpolation and the proposed interpolation method, the delay values giving zero time synchrony at the end of the frames were searched framewise. The search was performed with a resolution of 1/4 sample. The parameter σ_n was fixed to 172 samples, which is approximately two thirds of the frame length. The both resulting delay contours for a section of 10 frames are shown in Figure 3. The proposed method results in rapidly damping oscillation. In this example there are no oscillations after four frames. On the contrary, the conventional linear interpolation method results in a badly oscillating delay contour. These prolonged oscillations in the delay contour cause often annoying artifacts to the modified speech signal degrading the overall perceptual quality.

The delay selection produces as a byproduct additional information on the evolution of the signal. The value of the delay should not change significantly during purely voiced frame. Signal modification may impair the speech quality if the delay changes strongly. Thus it is reasonable to limit the allowed delay change. If the limits are exceeded, no modification is performed and the original signal is maintained.

4. MODIFICATION OF SIGNAL

As the delay contour has been selected, the signal modification itself can be performed. The shifts performed to the pitch cycle segments are selected by correlating segments in the weighted speech domain. However, the actual signal to be modified is the linear prediction residual signal. Each segment is allowed to be shifted $\pm\delta$ samples. For the first segment in frame the target signal $\tilde{w}(t)$ can be created recursively using

$$\begin{aligned} \tilde{w}(t) &= \hat{w}(t), & t \le t_{n-1} \\ \tilde{w}(t) &= \tilde{w}(t - d(t)), & t > t_{n-1} \end{aligned}, \qquad (3)$$

where $\hat{w}(t)$ is the synthesized weighted speech of previous frame. The shift which provides the highest correlation between the segment and the target signal is performed. The shift can be first searched with resolution of one sample and then the correlation values are upsampled to find the more precise shift. The target signal $\tilde{w}(t)$ is updated after each shift. Samples of $\tilde{w}(t)$ are first replaced with the modified weighted speech samples from the area of the shifted segment. Then, the samples following the updated segment are also updated using

$\widetilde{w}(t) = \widetilde{w}(t - d(t))$. These updates confirm higher correlation between successive pitch cycle segments in the modified speech signal (considering delay contour) and thus more efficient coding. The alternating segment shifting and target signal update procedures are continued until all segments in the frame have been processed.

The shifts of the first and the last segments in the frame are special cases which have to be performed particularly carefully. Before shifting the first segment, it has to be ensured that no high-energy regions exists in the residual signal close to the frame boundary, because shifting such a segment may cause artifacts. If there appears to be an important high energy region in the beginning of the first segment and notable shift is required for maximum correlation, it can be concluded that with provided initial delay value the frame is not suitable for signal modification, and the original speech frame is kept intact. The case of the last segment in the frame is quite similar. As was described in Section 3, the delay contour is selected such that no shifts is required for the last segment. However, because the target signal is repeatedly updated during signal modification, it is possible that some shifting is needed for the last segment. If the required shift is small (one sample or below at 12.8 kHz) and there is no high power region at the very end of the frame, the shift can be accepted. Otherwise the modification process is interrupted and the frame is coded in the full bit rate mode.

Shifting speech segments for maximum correlation gives final confirmation about the suitability of a frame for the signal modification. Normalized correlation values describe the similarity of successive pitch cycles. If values expose large dissimilarity, it can be concluded that the shape of the pitch cycle evolves rapidly and therefore the coding of speech frame at the low bit rate mode may result in poor speech quality. It is also useful to observe the sizes of shifts of successive segments. If the size of shifts differ noticeably it may degrade quality. It is reasonable to return back to the original signal and to the full bit rate mode if any of these measurements is not within allowed limits.

5. RESULTS

The classification following from the signal modification process works very reliably and confirms that only purely voiced frames are modified and coded in the low bit rate mode. The pitch cycle length of the modified frames typically evolve slowly and thus required modifications are usually minor. According to informal listening tests, without quantization the quality of modified speech remains transparent with the original signal. With the proposed delay contour shape, no quality degradations caused by oscillations were observed. The operation of the proposed signal modification method was also tested with quantization, modified frames were quantized with 7.6 kbps. Except for the adaptive codebook processing, the coding and quantization of the 7.6 kbps mode is performed equally as in 12.65 kbps full bit rate mode [4], only the number of bits which can be allocated for each parameter is smaller. The used bit allocation for voiced 7.6 kbps mode is presented in Table 1.

The number of modified frames can be controlled efficiently by setting limits for the normalized correlation values at the final stage of signal modification. Up to 45% of active speech can be modified if necessary. Informal listening tests with expert listeners indicate that while 10% of active speech is being modified and quantized at 7.6 kbps, the quality remains at same level with the 12.65 kbps full bit rate mode. While the number of modified frames is raised up to 30%, just a minor degradation in the overall quality can be perceived.

Table 1. Bit allocation for the voiced 7.6 kbps mode.

Parameter	Bits / Frame
LP Parameters	34
Pitch Delay	9
Pitch Filtering	4 (4×1)
Gains	24 (4×6)
Algebraic Codebook	80 (4×20)
Mode Bit	1
Total	152 bits = 7.6 kbps

6. CONCLUSIONS

A method for identifying and modifying purely voiced frames of wideband speech signal was presented. Signal modification is performed pitch and frame synchronously. It was shown that with a properly selected delay contour signal modification can be done frame synchronously without degrading the speech quality. Results indicate that proposed method implemented for the source-controlled ACELP speech coder provides good results at low bit rates. This shows also that the signal modification methods used in narrowband speech coding can be extended successfully also for wideband signals.

ACKNOLEDGEMENTS

M. Tammi wish to thank prof. Roch Lefebvre, prof. Jukka Saarinen and Mr. Vesa Ruoppila for making his one year visit to the University of Sherbrooke possible.

REFERENCES

[1] W.B. Kleijn, P. Kroon, and D. Nahumi, "The RCELP speech-coding algorithm," *European Transactions on Telecommunications*, Vol. 4, No. 5, pp. 573–582, 1994.

[2] W.B. Kleijn, R.P. Ramachandran, and P. Kroon, "Interpolation of the pitch-predictor parameters in analysis-by-synthesis speech coders," *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 1, pp. 42–54, 1994.

[3] Y. Gao, A. Benyassine, J. Thyssen, H. Su, and E. Shlomot, "EX-CELP: A speech coding paradigm", *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Salt Lake City, pp. 689–692, May 2001.

[4] B. Bessette *et al*., Techniques for high-quality ACELP coding of wideband speech," Eurospeech, Aalborg, Denmark, pp. 1997–2000, September 2001.

[5] 3GPP TS 26.190 "Adaptive multi-rate wideband speech transcoding," *3GPP Technical Specification.*