Exploiting Time Warping in AMR-NB and AMR-WB Speech Coders

Lasse Laaksonen[1], *Sakari Himanen*[2], *Ari Heikkinen*[2], *and Jani Nurminen*[2]

[1]Digital and Computer Systems Laboratory
Tampere University of Technology, Tampere, Finland
lasse.laaksonen@tut.fi

[2]Speech and Audio Systems Laboratory
Nokia Research Center, Tampere, Finland
sakari.himanen@nokia.com ari.p.heikkinen@nokia.com jani.k.nurminen@nokia.com

Abstract

In this paper, a time warping algorithm is implemented and its performance is evaluated in the context of Adaptive Multi-Rate (AMR) wideband (WB) and narrowband (NB) speech coders. The aim of time warping is to achieve bit savings in transmission of pitch information with no significant quality degradations. In the case of AMR-NB and AMR-WB speech coders, these bit savings are 0.65-1.15 kbit/s depending on the mode. The performance of the modified AMR speech coders is verified by subjective and objective measures in error-free conditions. MOS tests show that only slight, statistically insignificant degradation of speech quality is experienced when time warping is implemented.

1. Introduction

The Code Excited Linear Prediction (CELP) speech coding technique [1] has been very successful resulting in several CELP-based speech coding standards. This is true especially for coders employing the telephone bandwidth of 200-3400 Hz, but also for wider bandwidths. Recent examples of such coders are the AMR-NB and AMR-WB speech coders standardized by 3GPP, and 3GPP and ITU-T, respectively.

A very important extension of CELP is the relaxed waveform matching CELP (RCELP) [2] where, as the name implies, constraints to the waveform matching of analysis-by-synthesis scheme are relaxed aiming at reducing the bit rate and simultaneously maintaining high speech quality. In the method, the original speech signal is time warped in such a manner that its pitch period contour matches the synthetic pitch period contour. In the Enhanced Variable Rate Coder (EVRC) [3], the RCELP coding approach with time warping is exploited.

In this paper, the time warping algorithm of the EVRC speech coder is adopted to AMR-NB and AMR-WB speech coders, and its performance is evaluated by both subjective and objective measures.

In Section 2 of this paper, descriptions of both AMR-NB and AMR-WB speech coders are given. The details of the implemented time warping method are described in Section 3, whereas the performance of time warping in the AMR-NB and AMR-WB coders is discussed in Section 4. Finally, conclusions are given.

2. AMR-NB and AMR-WB speech coders

The AMR speech coders are based on the algebraic CELP (ACELP) technology [4] employing multiple coding modes. In these modes different bit rates are employed for speech and channel coding to optimize speech quality to the prevailing channel conditions. The bit allocations for both coders in each mode are depicted in Table 1. In the AMR-WB coder, the Voice Activity Detection (VAD) bit is omitted in Table 1.

2.1. AMR-NB

The AMR-NB speech coder [5] operates at eight different bit rates: 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.2, and 12.2 kbit/s. The 12.2 kbit/s mode is equivalent to the GSM Enhanced Full-Rate (EFR) [6] coder while the 7.40 kbit/s mode is equivalent to EFR coder for the IS-136 system [7].

The frame size of the AMR-NB speech coder is 20 ms divided into four subframes of equal length. In the 12.2 kbit/s mode, two sets of linear prediction (LP) parameters are computed and jointly quantized for each frame. For all the other modes, one set of LP parameters are computed and quantized for each frame.

From the weighted speech signal, an open-loop pitch estimate is computed once per frame in the two lowest modes and twice per frame in the other modes to restrict the range for the closed-loop search. The absolute pitch lag, or equivalently the adaptive codebook index, is transmitted for the first subframe in the two lowest modes while in the other modes absolute values are transmitted for the first and third subframes. For the other subframes differential coding is used. A fractional pitch lag is used in all modes, the exact precision depending on the mode.

In the algebraic codebook, the innovation vector contains non-zero pulses having the amplitudes +1 or –1. The bit rate used for the algebraic codebook ranges from 1.8 kbit/s to 7.0 kbit/s, thus defining the main differences between different modes. For adaptive and algebraic codebook gains, scalar quantization is employed in the 7.95 and 12.2 kbit/s modes, whereas joint quantization is used in the other modes.

For the 4.75, 5.15, 5.90, 6.70, and 10.2 kbit/s modes, the algebraic codebook gain is adaptively smoothed using the actual quantized gain and average gain values. To reduce the sparseness of the algebraic codebook excitation, an adaptive anti-sparseness postprocessing is employed in the 4.75, 5.15, 5.90, 6.70, and 7.95 kbit/s modes. An adaptive postfilter is used in each mode to enhance the perceptual quality of the coded speech.

Table 1: Bit allocations of AMR-NB and AMR-WB speech coders

AMR-NB			AMR-WB		
Mode	Param.	Bits	Mode	Param.	Bits
12.20 kbit/s	LP coeff.	38	23.85 kbit/s	LP coeff.	46
	ada. CB	30		ada. filt.	4
	alg. CB	140		ada. CB	30
	ada. gain	16		alg. CB	352
	alg. gain	20		gains	28
	–	–		HB gain.	16
10.20 kbit/s	LP coeff.	26	23.05 kbit/s	LP coeff.	46
	ada. CB	26		ada. filt.	4
	alg. CB	124		ada. CB	30
	gains	28		alg. CB	352
	–	–		gains	28
7.95 kbit/s	LP coeff.	27	19.85 kbit/s	LP coeff.	46
	ada. CB	28		ada. filt.	4
	alg. CB	68		ada. CB	30
	ada. gain	16		alg. CB	288
	alg. gain	20		gains	28
7.40 kbit/s	LP coeff.	26	18.25 kbit/s	LP coeff.	46
	ada. CB	26		ada. filt.	4
	alg. CB	68		ada. CB	30
	gains	28		alg. CB	256
	–	–		gains	28
6.70 kbit/s	LP coeff.	26	15.85 kbit/s	LP coeff.	46
	ada. CB	24		ada. filt.	4
	alg. CB	56		ada. CB	30
	gains	28		alg. CB	208
	–	–		gains	28
5.90 kbit/s	LP coeff.	26	14.25 kbit/s	LP coeff.	46
	ada. CB	24		ada. filt.	4
	alg. CB	44		ada. CB	30
	gains	24		alg. CB	176
	–	–		gains	28
5.15 kbit/s	LP coeff.	23	12.65 kbit/s	LP coeff.	46
	ada. CB	20		ada. filt.	4
	alg. CB	36		ada. CB	30
	gains	24		alg. CB	144
	–	–		gains	28
4.75 kbit/s	LP coeff.	23	8.85 kbit/s	LP coeff.	46
	ada. CB	20		ada. CB	26
	alg. CB	36		alg. CB	80
	gains	16		gains	24
–	–	–	6.60 kbit/s	LP coeff.	36
				ada. CB	23
				alg. CB	48
				gains	24

2.2. AMR-WB

The AMR-WB speech coder [8] consists of nine different modes: 6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, and 23.85 kbit/s. In the coder, the input signal is sampled at 16 kHz, but is downsampled to 12.8 kHz before encoding. At the decoder, the synthesized signal is upsampled back to 16 kHz, and the high frequencies are generated by feeding scaled (parameter 'HB gain' in Table 1) and bandlimited white noise through the modified LP filter.

The frame size of the AMR-WB speech coder is 20 ms divided into four 5 ms subframes. The LP analysis and quantization is performed once per frame. Compared to the AMR-NB coder, a modified filter more suitable for wideband signals is used to generate the weighted speech.

From the weighted speech, open-loop pitch analysis is performed once per frame in the 6.60 kbit/s mode and twice per frame otherwise. The absolute pitch lag is transmitted for the first subframe in the 6.60 kbit/s mode while in the other modes absolute values are transmitted for the first and third subframes. For the other subframes differential coding is used. As in the AMR-NB coder, a fractional pitch lag is used in all modes. In order to enhance the pitch prediction performance for wideband signals, a frequency dependent pitch predictor (parameter 'ada. filt.' in Table 1) is used in all modes excluding the two lowest modes. In this approach, either a low-pass filtered or conventionally generated codevector can be selected based on the error criterion. In the algebraic codebook, similar structure is used as in the AMR-NB speech coder. However, very large codebooks are used to guarantee high speech quality.

To reduce the sparseness of the algebraic codebook excitation, anti-sparseness postprocessing is employed in the 6.60 and 8.85 kbit/s modes. Also, adaptive postfiltering is used for these modes. In all modes, a nonlinear gain smoothing technique is applied to the algebraic codebook gain in order to enhance the excitation in noisy conditions. Finally, a pitch enhancement procedure is used to emphasize the higher frequencies of the total excitation more than the lower frequencies.

3. Time warping implementation

The implementation of time warping for the AMR-NB and AMR-WB speech coders is rather straightforward, since the required modifications affect almost solely the long-term prediction (LTP). Because in the RCELP coder it is assumed that the real pitch contour evolves smoothly enough, the pitch lag estimates between successive frames can be interpolated to construct a synthetic delay contour. Here, this approach leads to a significant decrease in the number of bits needed for pitch lag information in a frame, since only one pitch lag needs to be transmitted. According to Table 1, for each bit rate it should be possible to reduce the number of bits without notable degradations by 13, 13, 17, 17, 19, 21, 19, and 23 bits per frame in the AMR-NB coder. Equivalently, in the AMR-WB coder the savings would be 15, 18, 22, 22, 22, 22, 22, 22, and 22 bits for each of the nine bit rates. These savings convert to an average bit rate reduction of 0.65-1.15 kbit/s. Thus, only 7 bits in AMR-NB and 8 bits in AMR-WB are used to transmit an integer lag value.

To execute the time warps, more look-ahead is needed than there is available in the two AMR speech coders. Therefore, the look-ahead of both coders is increased to 10 ms, which is still suitable for real-time applications. In the modified AMR coders exploiting warping, an open-loop pitch lag estimate is used to construct the synthetic pitch contour on a frame-wise basis. The synthetic pitch contour is obtained by linearly interpolating between the current delay estimate and the delay estimate of the previous frame. The resulting pitch contour is thus piecewise linear and evolves smoothly.

The time warping method is based on the generalized analysis-by-synthesis paradigm, which essentially suggests that a modified speech signal resulting in the best coding performance should be selected for coding. In the modified AMR-NB and AMR-WB coders, a search is thus performed to find the optimal modified signal, which is then encoded utilizing the conventional AMR ACELP algorithm. However, the LTP is decoupled from the analysis-by-synthesis stage, since the pitch contour is set a priori.

In general, it is convenient to perform the signal modification on the LP residual rather than the speech signal itself to prevent moving the formant locations. The optimal modified signal is found by time-aligning the original input speech signal with the long-term predictor contribution, which is as yet unscaled. This is achieved through first determining a target signal for the modification by mapping the past modified LP residual to the synthetic pitch contour. The actual modification is then performed by determining the most suitable time shift for each of the shifting segments.

Time warps with resolutions of 1/8 samples or 1/6 for the AMR-NB and AMR-WB (at 12.8 kHz), respectively, were used to match the current LP residual segment with the target signal. The maximal shift allowed was 4 or 7 samples in relation to the previous segment for the two coders, respectively. However, in the case of low signal periodicity, the maximal shift was more severely restricted.

Great care should be exercised in choosing the shifting segment boundaries. Modifying the residual signal without degrading the perceptual quality of the synthesized speech signal is possible only if all the significant features in the residual signal are preserved. Since some parts of the residual signal may be omitted or repeated at the boundaries between shifting sections, no pitch pulses should be located in those regions. Thus, in order to safely complete the modification of the signal, the locations of the pitch pulses in the residual must be found.

As suggested above, the modification is carried out in short segments. Each segment optimally contains no more than one pitch pulse. Hence, the shifting segments, or LTP subframes, are not the same as the subframes used in calculating the fixed codebook innovation. However, the total excitation for a subframe can be calculated after the LTP contribution is computed. In the modified AMR coders, once each pitch pulse was found by locating a significant energy concentration in the LP residual, a safety margin of at least 0.625 ms was used before and after the pulse location to guarantee that the section boundary would fall in a low-energy region. Segments that do not contain a high-energy pulse should be shifted equally with the previous segment.

The time warping procedure creates asynchrony between the original and the reconstructed signals. Since the time-asynchrony tends to increase over time, the accumulated shift between the original and modified signals must be controlled. The drift can be controlled through the delay contour in a very intuitive way: if the modified speech is considerably lagging the original speech signal, the pitch delay estimate is increased; the opposite holds, when the original signal is too much behind the modified signal. However, when the periodicity is high, the pitch delay cannot be biased as biasing could degrade the speech quality.

In the AMR-NB and AMR-WB coders, 2.5 ms was used as a threshold value for the accumulated shift, after which it was evaluated whether to change the delay estimate value. However, this was only permitted, when the periodicity of the residual was low. During unvoiced segments the accumulated shift can generally be zeroed without degrading the perceived speech quality. This helps to control the shift quite effectively.

4. Performance analysis

To evaluate reliably the performance of the time warping implementation in both the AMR-NB and AMR-WB speech coders, listening tests were conducted. The selected testing methodology was Absolute Category Rating (ACR). The tests included six different AMR modes with conditions for both the original and time-warped coders. In addition to direct reference condition, two conditions corresponding to two different modes with time warping of the input speech signal without coding were also included. To ensure unbiased distribution of listener ratings, also five MNRU conditions were used. In the tests, clean speech with nominal level of –26 dBov was used.

The modes used with AMR-NB test were 4.75, 5.15, 5.90, 6.70, 7.40, 7.95 kbit/s. The warped signals with no quantization were derived from the 4.75 and 7.95 kbit/s modes. The size of the listening panel was 20 persons. The average Mean Opinion Scores (MOS) for each condition are shown on the left side of Table 2. The results show that the time-warped AMR-NB implementation scores consistently slightly below the corresponding original AMR-NB, but statistically (95% confidence level) the coders are equal except the 4.75 kbit/s mode. This was observed to be mainly due to joint quantization of adaptive and algebraic codebook gains in 1st and 2nd subframes and in 3rd and 4th subframes, respectively, which yields in sub-optimal evolvement of the target for the LP residual modification. The warped speech conditions scored equally to the direct references. No direct dependencies between the gender of the talker and the performance of the time warping implementation were found.

The modes used with AMR-WB test were 6.60, 8.85, 14.25, 15.85, 18.25, 19.85 kbit/s. The modes used with plain warping of input speech signal were 6.60 and 19.85 kbit/s. In this test, the size of the listening panel was 24 persons. The average MOS scores for each condition are shown on the right-hand side of Table 2. As with the case of AMR-NB coder, the time-warped AMR-WB versions were judged to be slightly worse than the original coders, but the difference is not statistically significant. However, the warped speech with no quantization scored more than 0.5 MOS lower than direct reference. This is most likely due to downsampling of input signal to 12.8 kHz. Similarly to the AMR-NB coder versions, no gender dependencies were found. However, both coders scored slightly lower with male talkers.

In the performance evaluation, objective measures were also used, including e.g. Perceptual Evaluation of Speech Quality (PESQ) [9] software tool. With both coders, the PESQ results correlate quite well with the listening test results.

Table 2: Listening test results

AMR-NB		AMR-WB	
Condition	Score	Condition	Score
MNRU 06 dB	1.21	MNRU 06 dB	1.14
MNRU 12 dB	1.59	MNRU 12 dB	1.42
MNRU 18 dB	2.40	MNRU 18 dB	1.99
MNRU 24 dB	3.31	MNRU 24 dB	2.50
MNRU 30 dB	3.56	MNRU 30 dB	3.21
Direct	3.81	Direct	4.42
Direct warp 5.15	3.88	Direct warp 6.60	3.67
Direct warp 7.95	3.85	Direct warp 19.85	3.74
4.75 kbit/s	3.15	6.60 kbit/s	2.32
4.75 kbit/s warp	2.48	6.60 kbit/s warp	2.19
5.15 kbit/s	3.06	8.85 kbit/s	3.00
5.15 kbit/s warp	3.05	8.85 kbit/s warp	2.61
5.90 kbit/s	3.35	14.25 kbit/s	3.55
5.90 kbit/s warp	3.18	14.25 kbit/s warp	3.44
6.70 kbit/s	3.39	15.85 kbit/s	3.55
6.70 kbit/s warp	3.13	15.85 kbit/s warp	3.31
7.40 kbit/s	3.48	18.25kbit/s	3.48
7.40 kbit/s warp	3.49	18.25 kbit/s warp	3.36
7.95. kbit/s	3.50	19.85 kbit/s	3.52
7.95 kbit/s warp	3.35	19.85 kbit/s warp	3.47

5. Conclusions

In this paper, a time warping algorithm was implemented and evaluated in the AMR-NB and AMR-WB speech coders. The adoption of the time warping technique resulted in bit savings of 0.65-1.15 kbit/s in transmission of pitch information, depending on the mode. The performance of the modified AMR speech coders was evaluated in error-free conditions using both subjective and objective measures. The time warping was observed to result in slight quality degradation of the reconstructed speech but the differences were not statistically significant for the majority of the modes.

6. References

[1] B.S. Atal and M.R. Schroeder, "Stochastic coding of speech signals at very low bit rates", *Proc. of IEEE International Conference on Communications*, pp. 1610-1613, 1984

[2] W.B. Kleijn et al., "The RCELP speech coding algorithm", *European Transactions on Telecommunications*, Vol. 5, No. 5, pp. 573-582, 1994.

[3] TIA/EIA/IS-127, "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems", *Telecommunications Industry Association Document*, February 1996.

[4] C. Laflamme et al., "16 kbps wideband speech coding technique based on algebraic CELP", *Proc. of Int. Conf. Acoust., Speech, and Signal Processing*, Toronto, Canada, pp. 13-16, 1991.

[5] E. Ekudden et al., "The adaptive multi-rate speech coder", *Proc. of IEEE Workshop on Speech Coding*, pp. 117-119, 1999.

[6] K. Järvinen et al., "GSM enhanced full rate speech codec", *Proc. of Int. Conf. Acoust., Speech, and Signal Processing*, Munich, Germany, pp. 771-774, 1997.

[7] T. Honkanen et al., "Enhanced full rate speech codec for IS-136 digital cellular system", *Proc. of Int. Conf. Acoust., Speech, and Signal Processing*, Munich, Germany, pp. 731-734, 1997.

[8] J. Rotola-Pukkila et al., "AMR wideband coder – leap in mobile communication voice quality, *Proc. of Eurospeech,* Aalborg, Denmark, pp. 2303-2306, 2001.

[9] ITU-T Rec. P.862. "Perceptual evaluation of speech quality (PESQ)", *ITU Recommendation*, 2001