Target Correlation Approach for Modification of Low Correlated Pitch Cycles of Residual Speech

HASSAN FARSI
Dept. of Electronic and Electrical Eng.,
Faculty of Eng., University of Birjand,
Birjand, Iran.

Abstract- Pre-assumption for low bit rate speech coding is that pitch evolves smoothly and pitch cycles are highly correlated for voiced speech. Due to non stationary characteristics of speech signal, this assumption is sometimes inaccurate and therefore the performance of speech coder is affected. In this paper, low correlated pitch cycles are specified and then modified using target correlation concept which maintaining perceptual speech quality identical to original speech. Since the modification is performed independently from speech coder, this can be applied as a pre-processing in low bit rate speech coding to provide more regular speech.

Keywords: Speech coding, Linear prediction, pitch cycle evolutions, pitch modification.

1 Introduction

Low bit rate speech coders rely on pitch and pitch cycles evolves smoothly during a frame. According to this assumption pitch, voicing and other speech model parameters are estimated. The accuracy of the estimated parameters play a major role in speech quality. For instance, *Waveform Interpolation (WI)* encoders require the pitch period at every extraction point in order to perform *Characteristic Waveform* (CW) extraction. In WI decoding, a pitch value at every sample point is required to construct a phase track, and then to convert a two-dimensional surface to a one-dimensional signal [1]. In *Mixed Excitation Linear Prediction* (MELP), the voicing level of speech is calculated by using the estimated pitch and the normalized autocorrelation value for five frequency bands [2].

Irregular pitch variations and low correlated pitch cycles can lead to inaccurate pitch and voicing level estimation. This affects the CW extraction in WI encoders and phase track construction in WI decoders. Moreover, in MELP, the voicing level of speech is affected by inaccurate pitch estimation, which degrades synthesised speech quality.

Pitch smoothing is applied as a pre-processing method, which modifies the residual signal such that a more smoothly evolving pitch contour is achieved [3]. In order to provide more regular speech, we specify and modify low correlated pitch cycles while maintaining speech quality identical to original speech.

This paper is organised as follows. In section 2, we present the effect of low correlated pitch cycles on pitch estimation and inaccurate decomposition of CW to Slowly Evolving Waveforms (SEW) and Rapidly Evolving waveforms which can affect the quality of synthesised speech. In section 3, the proposed modification is introduced. The results are analysed in section 4 and the conclusions are drawn in section 5.

2. Position of the problem

In order to find out the effect of low correlated pitch cycles on the estimated speech parameters, we apply a speech segment (Fig. 1) containing low correlated cycles to WI coder. The first step involves that the residual signal is obtained from the speech signal using the LP analysis filter. The corresponding residual signal is used for pitch estimation and the intermediate pitch values are obtained by interpolation of two adjacent pitch periods. Next, the CWs are extracted and aligned to construct a CW surface. The CW extraction is performed every 2 ms.

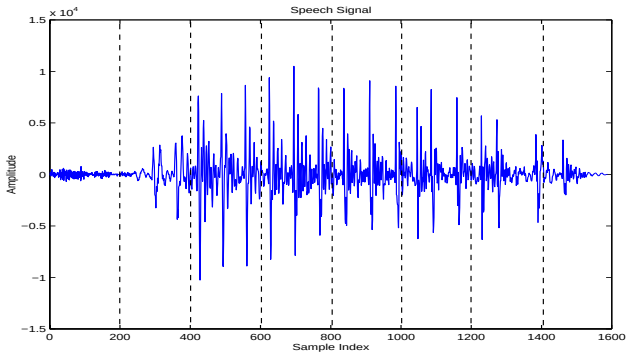

Fig. 1: Speech signal with irregular pitch variations and low correlated cycles. Dash lines show frame boundaries.

After the alignment and the normalisation procedures on the extracted CWs, the CW surface is constructed.

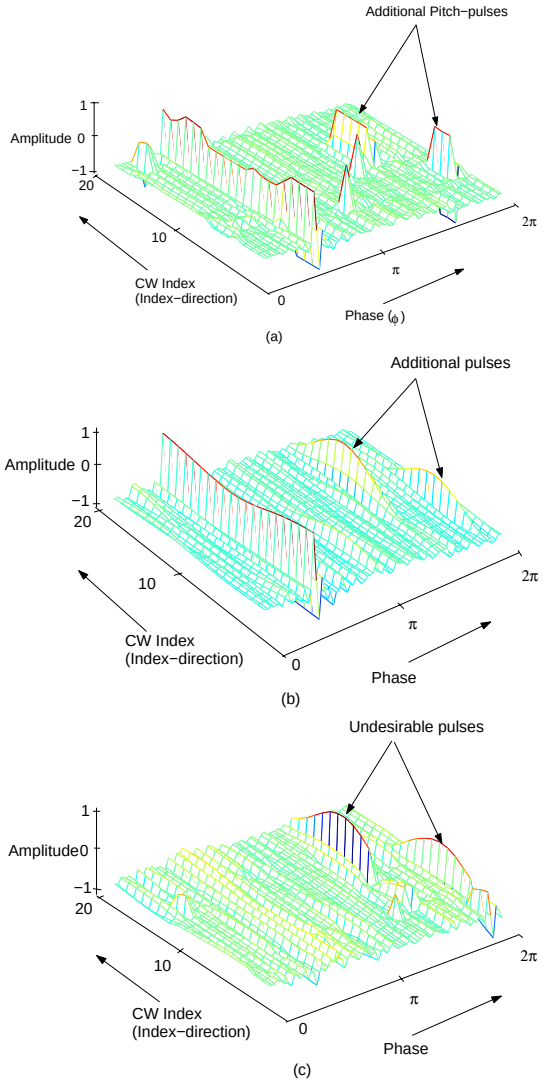

Fig.2: Resulting CW (a), SEW (b) and REW (c) for the sixth frame in Fig.1 containing low correlated cycles.

Fig. 2 shows the resulting CW surface for the sixth frame. For voiced speech, the resulting CW surface is expected to have maximum smoothness in CW-extraction time direction (index direction)[1]. However, due to low correlated cycles and corresponding CWs, some of the pitch pulses are not phase aligned. Next, the CW surface is decomposed into SEW and REW by low-pass and high-pass filtering with a cut-off frequency of 20Hz. It is expected that the voiced speech is transferred to the SEW surface, but the misalignment of the pitch pulses causes the bandwidth of the evolution spectrum to exceed 20Hz.As a result, a part of the voiced speech is transferred into the REW surface and the decomposition is performed incorrectly. In the next section a method, which slightly modifies the residual signal is introduced. This modification ensures that the pitch cycles evolve more smoothly during a frame and the speech model parameters are more accurately estimated.

3- Proposed Method

In the following sections, a pitch cycle is centred by a pitch pulse with length of the interpolated pitch value. The basic idea of the target correlation approach is to modify low correlated pitch cycles of the residual signal. Thus, it is required that the target contains highly correlated pitch cycles. The low correlated cycles are searched by computing the normalised cross-correlation between the pitch cycles of residual and the relative pitch cycles of the target signal and comparing it against a threshold (Fig. 3). These cycles are modified using the high correlated cycles (section 3.2). Thus, in the first step, the requirement is to construct the target signal.

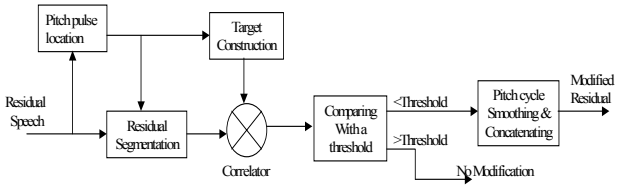

Fig. 3: A block diagram of target correlating based for pitch cycle evolutions smoothing.

3.1 Target construction

Ideally, the target signal should be as close as possible to the excitation of the vocal tract. Since the signal is unknown, the original residual signal can be used to design the target. The construction algorithm attempts to remove artefacts introduced by the standard LP method from the residual waveform. This is accomplished by guaranteeing a smooth evolution of pitch pulse shapes during voiced speech segments. The production of an individual target cycle is first described and is then followed with the algorithm that constructs a target frame.

3.1.1 Target cycle construction

In the first step, the pitch cycles are extracted using the estimated pitch value, P. Each pitch cycle contains

one centred pitch pulse to minimize boundaries energy. Consider L consecutive pitch cycles in the LP residual signal. After normalisation to unit energy and appropriate alignment we obtain the set of cycles $y_0, y_1, \ldots, y_{L-1}$.

The target cycle, x, is considered to have maximum correlation with vectors $y_0, y_1, \ldots, y_{L-1}$.

$$x = \arg\max_{\|x\|=1} \sum_i \left\| y_i^T x \right\|^2 \quad (1)$$

The operator $\|.\|^2$ denotes the 2-norm. i.e. $\|x\|^2 = x^T x$.

Equation 1 can be rewritten as:

$$\sum_i \left\| y_i^T x \right\|^2 = (Y^T x)^T (Y^T x) = x^T Y Y^T x \quad (2)$$

Thus x is obtained by:

$$\frac{\partial \Phi(x)}{\partial x} = 0, \qquad \Phi(x) = x^T Y Y^T x \quad (3)$$

Where the $P \times L$ matrix Y ($L < P$) is given by:

$$Y = [y_0 \; y_1 \ldots y_{L-1}] \quad (4)$$

The derivation vector of $\Phi(x)$ with respect to the vector x leads to:

$$2x^T Y Y^T = 0 \quad (5)$$

or equivalently:

$$Y Y^T . x = 0 \quad (6)$$

Since the rank of the $P \times P$ matrix YY^T cannot exceed the rank of Y or Y^T (which is L) [4] and $L < P$, YY^T is not a full rank matrix and therefore there is a non-zero solution. Equation 6 is solved by Singular Value Decomposition (SVD) method [4]. In this method, a matrix can be rewritten as the product of a column-orthogonal matrix **U,** a diagonal matrix **W** with positive or zero elements (the singular values), and the transpose of an orthogonal matrix **V.** Any column of **V** whose corresponding $\mathbf{w_j}$ of **W** is zero yields a solution. The vector, which maximizes Equation 2, is the target cycle x. Since the dimension of the matrix YY^T is $P \times P$, searching the target cycle x may require high computation, especially for the large pitch values. In order to reduce the cost of computation, we restrict the length of the cycles $y_0, y_1, \ldots, y_{L-1}$ to the minimum pitch value (2.5 ms equal to 20 samples for sampling frequency of 8 kHz). Since high-energy part of the cycles (which is pitch pulse region) has the main contribution in the cross-correlation value, the cycles $y_0, y_1, \ldots, y_{L-1}$ are centred at pitch pulse location with length of $P = 20$.

3.1.2 Target frame construction

The target frame is constructed through the target pitch cycles. Each pitch cycle is constructed using the procedure described in the previous section, considering the past target cycles, the current cycle, and possibly some cycles in the future. The current and the future cycles can be obtained from the original LP residual. The individual pitch cycles are extracted and normalised to have unity energy by using pitch pulse location information and the interpolated pitch values. They are then zero padded in time domain to have the same length and circularly aligned such that the cross-correlation between each cycle and the previous one is maximised. Each target cycle is constructed using the $\mathbf{n_1}$ target cycles from the past and the $\mathbf{n_2}$ current and future cycles. In other words, for any cycle y_l the relative target cycle is given by:

$$Y = [x_{-n_1+l} \cdots x_{l-1} \; y_l \cdots y_{n_2+l}]$$
$$x_l = \arg\max_{\|x_l\|=1} \left\| Y^T x_l \right\| \quad (7)$$

The resulting cycles are then rescaled and realigned with the original ones before replacing them in the residual waveform. Figure 4 illustrates the original residual and the constructed target signal for $\mathbf{n_1}$= $\mathbf{n_2}$=2. Compared to the original LP residual, the smooth evolution in the target cycles is clearly noticeable.

3.2 Pitch cycle evolution smoothing

After constructing the target signal, the normalised cross-correlation is computed between the cycles of the residual signal and the relative target cycles. Next, the ratio of the minimum to the maximum of the resulting values is computed. If the ratio is higher than a threshold (threshold = 0.85), no modification is performed.

Otherwise, the low-correlated cycles are replaced based on linear interpolation between the high-correlated cycles as given in Eq. 8.

$$y(i) = \frac{M+1-k}{M+1} C_{b0}(i) + \frac{k}{M+1} C_{b1}(i) \quad 1 \le k \le M, \; 0 \le i \le L$$
$$\tilde{C}_k(i) = \mu . y(i) \quad (8)$$

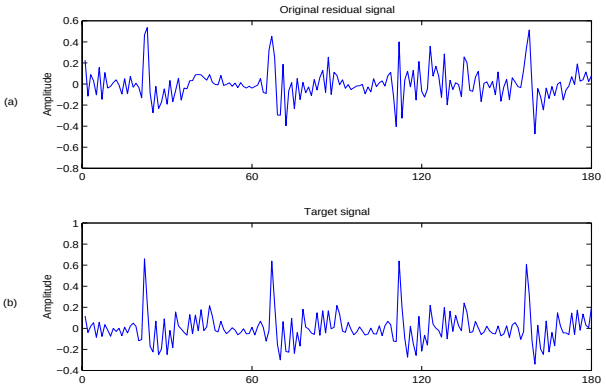

Fig. 4: (a) Original residual signal including non-smoothly pitch cycles evolutions, (b) the resulting target signal.

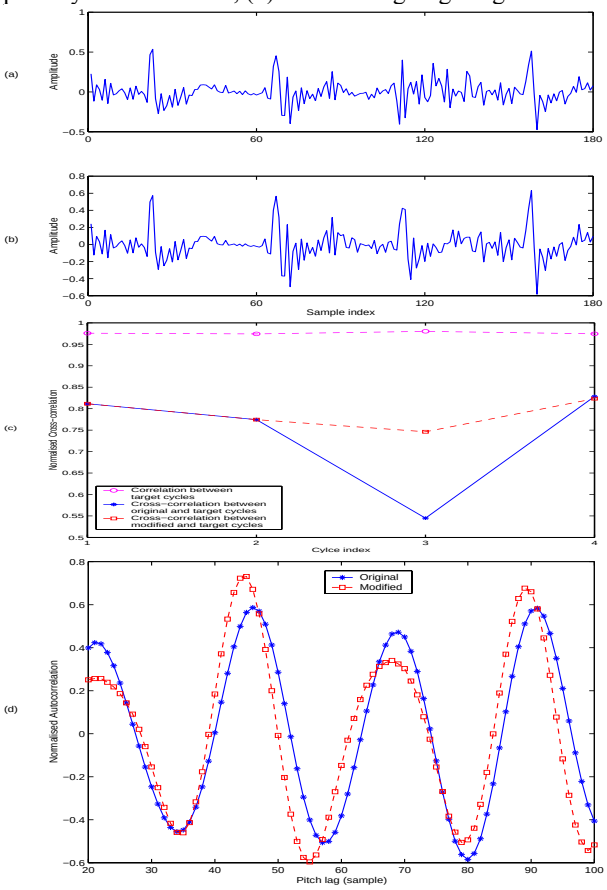

Fig. 5: (a, b) The original and modified residual signals, (c) the normalised cross-correlation between the target cycles and the original/modified residual cycles, (d) the normalised autocorrelation of the modified and original speech signals.

In this formula, C_{b0} and C_{b1} are the high correlated cycles before and after low correlated cycle C_k, M indicates the number of the low-correlated cycles placed between C_{b0} and C_{b1}. μ controls the resulting cycle energy to be identical to the original one and is given by Equation 9.

$$\mu = \sqrt{\frac{E_k}{E_y}} \tag{9}$$

Where E_k and E_y are the energy of the original cycle and the reshaped cycle y given in Eq. 8. In order to reduce the discontinuity energy, the overlap-and-add method is applied at the connection points. Fig. 5 shows the effect of the smoothing pitch cycles on the cross-correlation between successive residual cycles and also the normalised autocorrelation function. Obviously, due to higher correlation at a pitch lag of 45 samples and lower correlation at other pitch lags, especially at a pitch lag of 68 samples, a reliable pitch can be estimated.

4. The Proposed Method Evaluation

As detailed in the last section, the proposed method provides more regular speech such that the pitch can be more accurately estimated. This can be observed in Fig. 5-d. Obviously, due to higher correlation at a pitch lag of 45 samples and lower correlation at other pitch lags, especially at a pitch lag of 68 samples, a reliable pitch can be estimated. In [5] it is shown that more accurate estimated pitch leads to higher Pitch Prediction Gain (PPG), α, given by: $\alpha = \frac{R(T)}{R(0)}$,

where T is the estimated pitch and R(T) refers to the autocorrelation of the speech signal with time lag of T samples. The PPG can be used as a measure to indicate the accuracy of the estimated pitch. We therefore compute the PPG for the original and modified speech. If the difference between the PPG of the original frame and the relative modified one is less than a set threshold, we assume there is no preference between the estimated pitch values of the original and modified frames. Otherwise the estimated pitch value corresponding to the higher PPG is considered as the correct pitch value.

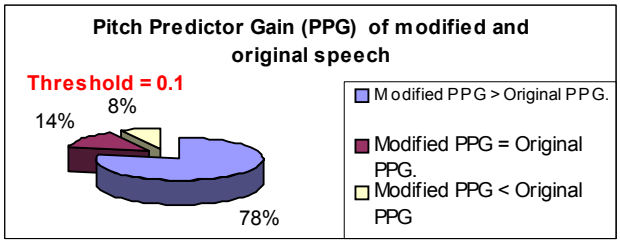

Fig.6: The accuracy of the estimated pitch values of the modified speech compared to original using PPG.

Figure 6 indicates the results of comparison of the PPGs of the original and the modified speech for two different thresholds 0.1.

The proposed method also affects the accuracy of the other parameters estimation, e.g., voicing level. In the following, we show that the pre-processor can also provide more accurate voicing level estimation. We apply both the original and the modified speech separately as input to the standard MELP 2.4 Kb/s. Figure 7 shows the original and the modified speech signals and the corresponding normalised autocorrelation function computed for pitch lags between 20-160 samples. The voicing decisions of five bands are computed by comparing the normalised autocorrelation value for the estimated pitch with a voicing threshold. It is observed that in spite of being strongly voiced speech (Fig. 5-a), only the first band of the original speech is estimated as voiced, whereas the first four bands of the modified are estimated as voiced.

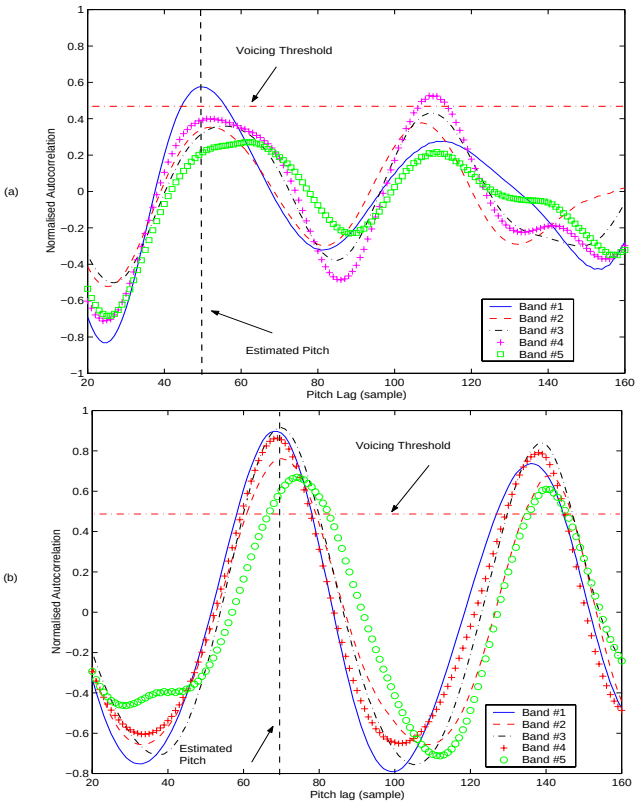

Fig. 7: The corresponding normalised autocorrelation function of the original (a), and the modified (b) first speech frame shown in Fig. 5(a,b), for five frequency bands.

In order to show the effect of the proposed method in frequency domain, the ratio of the Synthetic Spectral Matching of the Modified to the Original (SSMMO) speech was used as a measure. This is measured for original frames including irregular pitch variations and the corresponding modified frames. As an example, Fig. 8 shows the original and the modified speech signals and the corresponding synthetic spectral matching distortions. In Fig. 8-c, it is observed that for the frames including irregular pitch evolutions (sample index between 750-1700), the resulting distortion is significantly higher for the original compared to the modified one and consequently the SSMMO decreases (Fig.8-d).

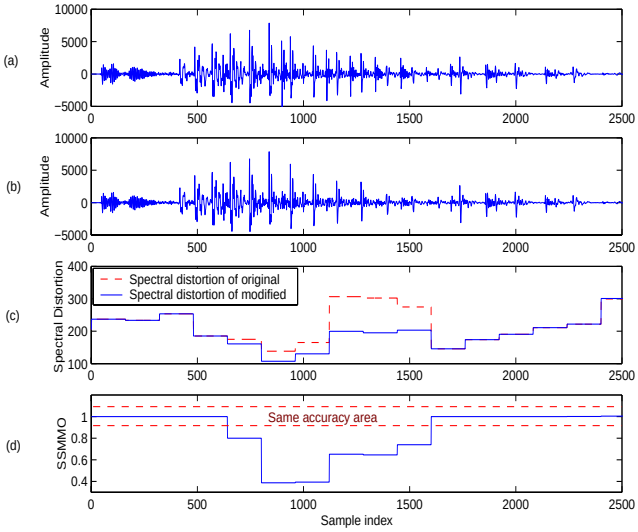

Fig. 8: (a) The original speech, (b)The modified speech, (c) The corresponding synthetic spectral matching distortion of the original and the modified, (d) SSMMO function with threshold TH=0.1.

When SSMMO<1, this indicates the estimated pitch of the modified speech is more accurate than the original one and vice versa. This measure is only computed for frames in which the proposed method is activated. Next, the SSMMO is calculated for the last four bands and also for the full band. Fig. 9 shows the accuracy of the voicing decision of the modified speech in comparison with the original one by using the spectral distortion with TH=0.1.

In the next experiment, the effect of the proposed method on the frame-to-frame pitch variations is presented. Since the modification is based on pitch cycle evolutions smoothing, it is expected that the pitch evolves smoothly from frame-to-frame as compared to the original counterpart. As an example, Fig. 10 shows the pitch variations of the original and the modified male and female speech signals. The circled segments show where the estimated pitch

values of the original speech result in high variations whereas the modified ones lead to smooth evolutions.

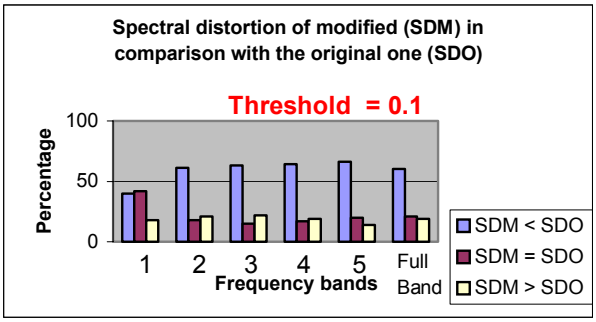

Fig. 9: Accuracy of the estimated voicing level using the spectral distortion.

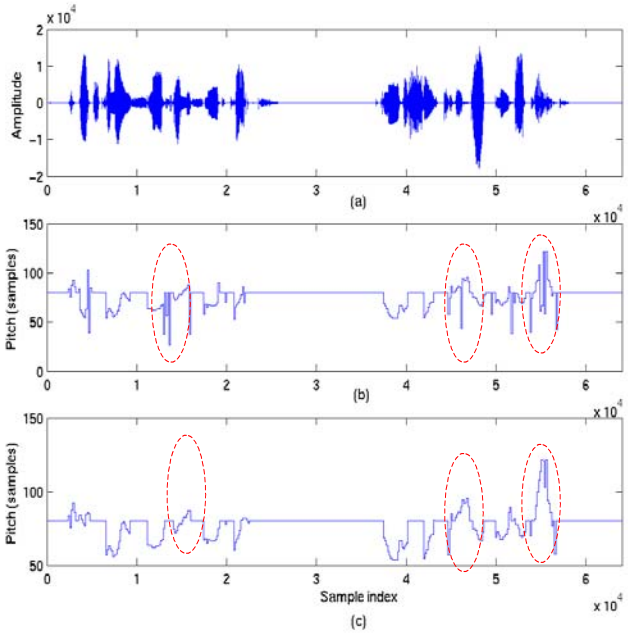

Fig. 10: (a) The original speech, (b) and (c), the estimated pitch values of the original and modified male speech, respectively.

For testing purposes, the proposed method was applied on short sentences and words (for both male and female) for which the modification was performed. An AB-test with 15 (4 trained and 11 untrained) listeners was carried out on the original and the processed speech files (including 8 speech sentences, 4 male and 4 female, and 10 single words) to investigate the perceptual speech quality of the modified before applying to a speech coder. In the next experiment, the effect of the proposed method on a speech coder was evaluated. This was performed by applying the original and the modified speech as the inputs to standard WI 2.8 kbps [1]. An A-B comparison test was carried out on the synthesised speech files. The results are shown in Table 1. The results obtained indicate that there is no statistical difference in perceptual quality between the original and the modified speech. However, the proposed method as a pre-processor in combination with the WI provides significantly better perceptual speech quality than the WI alone.

Table 1: Modified speech vs. original one and WI + proposed method vs. WI alone.

Speech Type	Better	Slightly better	Same	Slightly worse	Worse
Modified speech vs. Original	0.2	11.2	76.3	12.3	0
Proposed method + WI vs. WI	4.2	51.4	30.1	12.2	2.1

5. Conclusion

In this paper, the shortcomings of the low correlated pitch cycles on estimation of the speech model parameters were demonstrated. In order to overcome to these problems, a novel method based on target correlation concept was proposed. This method specifies and modifies the low correlated cycles and results in more regular speech which provides more accurate estimation of speech model parameters. The subjective listening tests show the quality of the modified speech is maintained identical to original speech. Moreover, applying the proposed method as a pre-processor in combination with WI 2.8 kb/s speech coder provides significantly better quality than the WI alone.

References:

[1] W. B. Kleijn, “Methods for waveform interpolation in speech coding," Digital Signal Processing, vol. 1, pp. 215-230, Jan. 1991.

[2] L.M. Supplee, R.P. Cohn, J. S. Collura, A.V. McCree, "MELP: The New Federal Standard at 2400 bps," *ICASSP97, Munich*, Germany.

[3] H. Farsi, A. kondoz,” Pre-processing method for pitch smoothing,” *IEE Electronics Letters,* Vol. 37, No. 21, 11 Oct. 2001, pp. 1314-1316.

[4] Franz E. Hohn. Elementary matrix algebra. Macmillan company, New York, 1973.

[5] A. M. Kondoz. Digital speech: coding for low bit rate communication systems. Third Edition, John Wiley, Chichester, UK, 2005.