A Pre-processing Method to Modify Irregular Pitch Variations for Quality Enhancement of Synthesised Speech

Hassan Farsi
Dept. of Electrical and Computer Eng., University of Birjand, Birjand, Iran
hfarsi@birjand.ac.ir

ABSTRACT

In low bit rate speech coders, pitch and voicing level estimation play an important role in quality of the synthesised speech. Although pitch usually evolves smoothly, sometimes it has irregular variations and as a result the estimated pitch and the voicing level differ from the real ones. This affects the performance of the speech coder.

We propose to use a new modification as a pre-processor. This methodology modifies the residual speech signal such that the pitch period evolves more smoothly without distorting perceptual speech quality. Thus, the pitch and the voicing level can be determined correctly. Experimental results show that combination of the proposed method with 2.4 Kb/s MELP coder provides better quality.

Keywords: MELP, Pitch smoothing, WI coder, pitch cycle, speech pre-processor.

1. INTRODUCTION

Most low bit rate speech coders rely on the pitch estimate in determination of other speech model parameters [1][2]. For instance, in a MELP coder, voicing strengths are estimated using the normalised autocorrelation values computed at the estimated pitch lag [14]. In a sinusoidal coder, the speech waveform is represented by the sum of sine waves; the frequency of sinusoids is computed using the estimated pitch value. In a WI coder, the estimated pitch value is used to determine the length of the CWs during the CW extraction procedure in the analysis stage and to construct a phase track, which is used to convert a two-dimensional surface to a one-dimensional signal during synthesis stage [15].

The basic assumptions for pitch estimation algorithms are:

- Voiced speech samples are correlated at specific time intervals called pitch period. Although usually these samples are usually highly correlated, they sometimes have low correlation and as a result the resulting normalised autocorrelation is unreliable for pitch estimation.
- Pitch evolves smoothly during a voiced frame. Although this usually happens, pitch occasionally has irregular variations, which can lead to inaccurate pitch estimation. The irregular pitch variations usually happen in transition frames where speech characteristics change from quasi-periodic to random like signal or vice versa and therefore the long-term correlation of speech signal is affected.

In order to overcome these problems, a methodology that provides more accurate estimation of the parameters is required. This methodology can be performed either inside a speech coder using alternative algorithms or outside as a pre-processor. In the first case, no modification is performed on the speech signal and alternative factors such as history and future of a parameter, spectral matching and etc. are used to find a better estimation. Whereas making more regular speech is the objective of the second method. This enables the algorithms used in speech coder to a more accurate pitch. This technique is called speech pre-processing since the speech signal is modified before passing to speech coder.

This paper is organised as follows: In the next section the existing pre-processors and techniques for pitch modification are presented. In section 3, we propose and describe a new pre-processor, which leads to smooth pitch evolutions and provides more regular speech such that the required parameters can be effectively estimated in a speech encoder. In
section 4, the effect of the new pre-processor is evaluated in i) pitch estimation ii) voicing level estimation iii) subjective listening tests. This is performed by comparing the case which the pre-processor is in combination with the standard MELP 2.4 Kb/s and the MELP alone. Finally, the paper ends with section 5 where the conclusions are drawn.

2. EXISTING PRE_PROCESSORS AND PITCH MODIFICATION TECHNIQUES

In the following subsections we address current pre-processors employed by speech coders and current techniques for pitch modification.

2.1 Existing Pre-processors

Existing pre-processors have been designed for special coders. For instance, in [3], Kleijn introduces a pre-processor in combination with a block-DFT based WI coder. This coding structure maintains the advantages of earlier WI coders and adds the asymptotically perfect reconstruction property. Since the alignment procedure employed by a WI encoder causes the relative phase loss of the CW, the WI coder does not give perfect reconstruction [3]. The alignment procedure is included in as a part of the pre-processor performed outside the WI coder. This is performed by moving a pitch pulse to the centre of a CW such that the modified segment is maximally correlated with the previous cycle and thus the employed alignment procedure in earlier WI coders is not required. The pre-processor employed in [4] performs high-pass and adaptive noise suppression filtering before estimation of speech parameters. After speech parameter estimation, the residual signal is modified to generate a target residual for the fixed codebook search in a CELP coder. A shifted target residual is generated using the past-modified residual and the delay contour of the current frame. This shifted residual is used as a target for shifting the residual of the current subframe. All pitch pulses in the original residual are shifted individually to match the delay contour of the modified target residual.

2.2 Existing Pitch Modification Techniques

Existing pitch-modification techniques can be classified into two groups. The aim of the first group is rendering speech at an arbitrary rate different from the original rate. This can increase the ease-of-use and the efficiency of speech reproduction equipment. A number of algorithms for high-quality time and pitch scaling have been reported in [5]. In fact, the common character of these algorithms is to change the duration of all pitch cycles by an arbitrary factor. In other words, all pitch cycles are scaled in the same way and thus in the existing techniques the irregular pitch variations present in the modified speech. In the second group, the alternative techniques are applied to the original speech before further parameter analysis is performed. For instance, in [6][7], a time warper is applied to the original speech to enhance the stationary of voiced speech segments. The main feature of the employed time warper is to remove the part of the frequency variation, which progresses linearly with time, without changing the time duration of that segment. As a result, irregular pitch-pulse locations remain in the modified speech and time-domain algorithms employed for pitch and voicing level estimation may fail. In addition, the quality of the modified speech is not necessarily maintained perceptually.

We propose a new pre-processor, which enhances the regularity of speech signal and also maintains the perceptual speech quality and thus can be used in combination with any speech coder.

3. PROPOSED PRE-PROCESSOR DESCRIPTION

The proposed pre-processing algorithm modifies the residual signal such that it is more convenient for coding. During the modification, the lengths of the pitch cycles in a frame are altered so as to evolve more smoothly. The modification is based on the local pitch estimation. The local pitch values are only used in the encoder and only the refined pitch value is transmitted to the decoder once per frame. The input to the pre-processor
is the linear prediction residual of the speech signal and an associated pitch track. The pitch period is estimated once in a frame using conventional autocorrelation based methods and the resulting estimate is then linearly interpolated for each pitch cycle. The output is a modified linear prediction residual, which is constructed by concatenation of modified/unmodified pitch cycles of the residual signal.

If the frame is unvoiced, no modifications are performed to the pitch value. During voiced sections the main task of the pre-processor is to smooth the pitch values of the pitch cycles while keeping long-term correlation of the speech signal and also maintaining the perceptual speech quality identical to the original. In the following sections, the operation of the pre-processor on a step-by-step basis is discussed.

3.1 Local Pitch Estimation

A simple method based on the energy of the LPC residual is employed by the Telecommunication Industry Association (TIA) Enhanced Variable Rate Coder (EVRC) [4], to detect the pitch pulses. The EVRC computes the pitch pulse locations by searching for a maximum in a five-sample moving energy window within a region one and half times the pitch period, and then finds the rest of the pitch pulses by searching recursively at a separation of one pitch period. In [8], the performance of the residual energy based pitch pulse location is improved using the Hilbert Envelope of Windowed LP Residual (HEWLPR). A robust pitch pulse detection algorithm based on the group delay of the phase spectrum has also been reported [9].

The proposed pre-processor requires a pitch pulse detection algorithm, which can detect the pulses at stationary voiced segments with high accuracy. Therefore, an improved pitch pulse detection algorithm as the algorithms used in the EVRC energy based pitch pulse detection and the HEWLPR is proposed for the pre-processor. In fact, the EVRC and the HEWLPR algorithms compute all possible pitch pulse locations and if there is a difference between the determined locations, the corresponding pitch pulses are modified during the first stage of pitch modification (section 3.2). Thus, after applying these two algorithms, two vectors are created. One containing the valid pitch pulse locations, \( V_{val} \), and the other one having the invalid pitch pulse locations, \( V_{inv} \). These vectors may be updated during the pitch contour construction. However, pitch pulse refinement (section 3.2) and pitch cycle modification (section 3.3) are performed using \( V_{inv} \) and \( V_{val} \) information.

In [10], it is proposed to use localised energy and adaptive threshold for pitch pulse locations. Initially all the possible pitch pulse locations are determined by using the localised energy of the residual signal, \( r(n) \), and an adaptive threshold, \( t(n) \). The localised energy, \( e(n) \), is computed by moving a rectangular window with length of five samples across \( r(n) \), and is given by equation 1:

\[
 e(n) = \frac{1}{5} \sum_{i=-2}^{2} |r(n+i)| \quad 2 \leq n < N - 2 \quad (1)
\]

Where \( N \) is sum of the current frame length, which is 180 samples and the buffered samples of the previous and the next frame. The length of this buffer is considered to be 20 samples. The adaptive threshold function \( t(n) \) is updated for each half pitch period by taking 0.65 of the maximum of \( e(n) \) in the pitch period symmetrically centred around the half pitch period chosen to calculate \( t(n) \), and \( t(n) \) is given by equation 2.

\[
 t(n_k - T_{1/4} + n_{T/2}) = 0.65 \max \left[ e(n_k - T_{1/4} + n_T) \right] \\
 \text{for } 0 \leq n_T < T_1 \text{ and } 0 \leq n_{T/2} < T_{1/2} \quad (2)
\]

Where

\[
 T_1 = \left\lfloor \frac{T + \frac{1}{2}}{2} \right\rfloor , \quad T_{1/4} = \left\lfloor \frac{T + \frac{1}{4}}{2} \right\rfloor , \quad T_{1/2} = \left\lfloor \frac{T + \frac{1}{2}}{2} \right\rfloor , \quad n_k = kT_{1/2} \\text{ for } 1 \leq k \leq \left\lfloor \frac{2N}{T} \right\rfloor , \text{ and } T \text{ is the pitch period.}
\]

For frame boundaries, the previous and next frame samples are used for localised energy.
The initial amounts of the threshold function, \( t(n) \), is given by equation 3.

\[
t(m) = 0.65 \max [e(T_{m/2})] \quad \text{for } 0 \leq m < T_{1/4} \tag{3}
\]

The samples locations, for where \( e(n) > t(n) \), are considered as regions which may contain pitch pulses if these regions are part of the current frame. If \( e(n) > t(n) \), for more than eight consecutive samples, those regions are ignored, since in those regions the residual energy is smeared, which is not a feature of pitch pulses. The maximum amplitude of each remaining region is considered as a possible pitch pulse location. If any of the two candidate locations are closer than 16 samples (which is considered the minimum pitch period) from each other only one, which has the higher local energy, is taken and the other one is considered as invalid pitch pulse and so a member of vector \( V_{\text{inv}} \).

Applying an adaptive threshold to estimate the pitch pulse locations from the localised energy is advantageous, especially for the segments where the energy of the LPC residual varies rapidly. Figure 1 shows this event for two occasions, (a) a male offset and (b) a female onset. The male speech frame has a pitch period of about 76 samples and two high-energy invalid pitch pulses. The female speech frame has a pitch of about 78 samples, which also consist of one high-energy invalid pulse. The energy function \( e(n) \) and the threshold function \( t(n) \) are also depicted in Figure 1, both \( e(n) \) and \( t(n) \) are shifted upwards for clarity. The figures also show that \( e(n) \) at the invalid pulses may be higher than \( e(n) \) at the valid pitch pulses. Therefore selecting the highest \( e(n) \) to detect a pitch pulse location as in [7] may lead to errors. However, since \( e(n) > t(n) \), for some of the irregular pulses as well as for valid pitch pulses, further refinements are required.

Due to reducing the competitions, the refinements are applied if the standard deviation of the vector including the estimated pitch-pulse locations is more than 10% of the vector mean.

The proposed pitch pulse locations refinement in [10] relies on the accuracy of the estimated pitch value, but in the case of irregular pitch variations, the estimated pitch value can be incorrect. Therefore, in order to obtain the valid pitch pulse locations two-stage refinement is proposed.

![Figure 1](image1.png)

**Figure 1:** (a) Male offset speech frame including two high-energy irregular pitch pulses, (b) Female onset speech frame including one high-energy irregular pitch pulse.

### 3.1.1 Pitch-pulse Location Refinement: First Stage

As shown in the previous section, both the valid and invalid pitch pulses are detected using concentrated energy measure and
The HEWLPR with adaptive threshold is used to separate the invalid pitch pulses from the valid ones. In [8], it is assumed that the speech signal within a pitch period is induced by a pulse at one epoch or an event. This epoch is defined as a representation of the Glottal Closure Instant (GCI), because the GCI induces the sound vibration and introduces most of the energy within each pitch period. An epoch occurs when the conditional probability density, or likelihood function, of the epoch is maximised. This can also be performed through the maximisation of a function called Maximum-Likelihood Epoch Determination (MLED) signal which is defined by equation 4 [8].

\[ \hat{f}(n_o) = \sum_{n=0}^{L-1} s(n + n_o) \hat{s}(n) \quad \text{for} \quad 0 \leq n_o < N \]  

Where \( N \) and \( L \) are the frame size and the length of the wavelet due to an epoch, \( \hat{s}(n) \).

Assuming that speech production can be modelled as an all-pole linear system, \( z \)-transform of the wavelet due to an epoch can be expressed by the equation 5.

\[ \hat{S}(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \]  

Where \( a_i \) and \( p \) are the coefficients and the order of the polynomial respectively. The MLED creates not only a strong and sharp epoch pulse, but also a set of weaker pulses representing the suboptimal epoch candidates within a pitch period. The energy ratio between the valid epoch pulse and the sub-pulses varies largely which results in ambiguity for pitch pulse location determination. This was overcome using a selection signal, \( g(n) \), superimposed on the GCI’s:

\[ g(n_o) = \left[ \hat{f}^2(n_o) + \hat{f}_H^2(n_o) \right]^{1/2} \]  

Where \( \hat{f}_H(.) \) is the Hilbert transform of \( \hat{f}(.) \) which can be identified as a filter with the transfer function:

\[ H(\omega) = \begin{cases} 
-j & 0 < \omega < \pi \\
0 & \omega = 0, \pi \\
-j & -\pi < \omega < 0 
\end{cases} \]  

Finally, the GCI Determination Signal (GCIDS) is the MLED signal \( \hat{f}(n_o) \) multiplied by the selection signal \( \hat{g}(n_o) \):

\[ \theta(n_o) = \hat{f}(n_o) \cdot \hat{g}(n_o) \]  

Where

\[ \hat{g}(n_o) = \begin{cases} 
g(n_o) - \bar{g}(n_o) & \text{if } g(n_o) \geq \bar{g}(n_o) \\
0 & \text{if } g(n_o) < \bar{g}(n_o) 
\end{cases} \]  

and

\[ \bar{g}(n_o) = \frac{1}{N} \sum_{n=0}^{N-1} g(n_o) \]  

The GCIDS compared with the MLED provides stronger and sharper epoch pulse and weaker sub-pulses, which is more suitable for more pitch pulse location determination. The following steps are performed to the GCIDS.

1. \( \hat{s}(n) \) is computed using 4ms of an impulse-response of the LP inverse filter.
2. The MLED signal is calculated through Eq. (4).
3. The Hilbert transform of the selection signal is computed using 256-points FFT of the MLED signal multiplied by \( H(\omega) \) and then taking the resulting FFT.
4. The selection signal is obtained using Eq. (6) and subtracted from its mean. The resulting signal is rectified so that the positive values are maintained.
5. The GCIDS is computed by multiplication of the MLED and the rectified selection signals.
Fig. 2 shows the residual signal and the corresponding GCIDS in comparison with localised energy variations. The experiments show that the valid pitch pulse locations appear in the GCIDS and the invalid pitch pulse locations are mostly removed and in few cases still appear with very low energy. Thus, the searching pitch-pulse procedure using the adaptive threshold is performed on the GCIDS signal. Consider \( n_p \) to be the pitch-pulse location found by the algorithm described in section 3.1. The maximum energy of GCIDS at interval \([n_p-2, n_p+2]\) is compared with the corresponding threshold. If it is higher than 95 percent of the threshold, the found pitch-pulse at \( n_p \) is considered as a valid pitch pulse. Otherwise, this is considered as an invalid pitch pulse and is refined by procedure described in section 3.2.

Assuming \( V_{evrc} \) and \( V_{hewlpr} \) are the vectors which include pitch pulse locations obtained in section 3.1 and the HEWLPR refinement, we compute vectors \( V_{val} \) and \( V_{inv} \) as follows:

\[
V_{val} = V_{evrc} \cap V_{hewlpr} \\
V_{inv} = V_{evrc} - V_{hewlpr}
\]

(11)

Next, the difference between each two successive pitch pulse locations is computed using \( V_{val} \) information. If the standard deviation of the resulting vector is more than 10% of its mean, the second stage of refinement based on the pitch contour is performed.

3.1.2 Pitch Contour Construction: Second Stage of Pitch Pulse Locations Refinement

As pitch has irregular variations, the pitch contour based on the estimated pitch values can be inaccurate. Therefore we construct the pitch contour based on the local pitch values in \( V_{val} \). The aim of the pitch contour construction is to change the irregular pitch values such that the pitch evolves smoothly during a frame. This is performed using the history and the future of the pitch variations. In order to have the history and the future of the pitch variation, we keep the local pitch information of the previous frame and also estimate the valid local pitch values of the next frame. If \( V_{val-pre} \), \( V_{val-cur} \) and \( V_{val-next} \) are the valid pitch pulse locations of the previous, current, and next frame respectively, the vector \( V_{valid} \) is defined by equation 12.

\[
V_{valid}(k) = \begin{cases} 
V_{val-pre}(k-N) & 1 \leq k \leq L_{val-pre} \\
V_{val-cur}(k-L_{val-pre}) & 1 \leq k \leq L_{val-cur} \leq L_{val-cur} + N \\
V_{val-next}(k-L_{val-pre}-L_{val-cur}) & 1 \leq k \leq L_{val-pre}-L_{val-cur} \leq L_{val-next} 
\end{cases}
\]

(12)

Where \( L_{val-pre} \), \( L_{val-cur} \) and \( L_{val-next} \) are the dimension of the vectors \( V_{val-pre} \), \( V_{val-cur} \) and \( V_{val-next} \), respectively and \( N \) is the frame size. There are two exceptions for onset and offset frames. Due to the nonexistence of the previous and the next pitch pulse locations for onset and offset frames, respectively, the locations of two next and two previous frames are used to construct \( V_{valid} \). Next, the difference between each two successive elements of the vector \( V_{valid} \) is calculated. The resulting vector, \( V_{local} \), indicates the local pitch variations where each local pitch cycle is started and ended by two successive pitch pulses. The irregular pitch
values appear as maximum or minimum values in the vector \( V_{\text{local}} \). Figure 3 shows a residual signal with the corresponding vector \( V_{\text{local}} \).

The pitch contour is constructed by regular pitch values and therefore it is required that these are separated from irregular values. The separation is performed based on the pitch average. In order to obtain the pitch average, the maximum and minimum values (if existing) are ignored in calculation of the mean of the vector \( V_{\text{local}} \). The maximum and minimum values are defined as the points that have more than 10% variation with their neighbouring points. If the vector length resulting from ignoring the maximum and minimum values is more than four, searching and ignoring these points is performed again. Next, the pitch average for each frame is obtained by calculating the mean of the resultant vector \( V_{\text{local}} \). If there is not enough information (at least two points required) to calculate the pitch average of the current frame, \( \bar{P}_{\text{avg-car}} \), it is obtained by the arithmetic mean of the previous and the next pitch average. The regular and irregular pitch values of the current frame are evaluated by the measure given in Eq. 13.

\[
R(k) = \frac{V_{\text{local}}(k) - \bar{P}_{\text{avg-car}}}{\bar{P}_{\text{avg-car}}} \quad L_1 \leq k \leq L_2 \quad (13)
\]

Where \( L_1 \) and \( L_2 \) select the elements of \( V_{\text{local}} \) within the current frame. If \( R(k) \) is less than 10%, the corresponding local pitch is considered as a regular pitch value. Otherwise, pitch doubling and pitch halving is checked. In order to prevent pitch doubling, the relative \( R(k) \) is checked. If it is between 95 - 105% there is a possibility of pitch doubling. This may happen when a low-energy pitch pulse is placed between two high-energy pitch pulses. Thus, maximum local energy of samples placed between 45-55% of the relative local pitch cycle is compared against the energies of both sides of a pitch pulse. If it is more than 50% of the minimum pitch-pulse energies, we consider it as being a pitch pulse in this area and therefore the vectors \( V_{\text{val}} \) and \( V_{\text{local}} \) are updated. Half-pitch errors may occur if the relative \( R(k) \) is between 45-55%. This may happen when a high-energy-irregular pitch pulse is placed between two regular pitch pulses and appeared in GICIDS. In order to avoid such half pitch errors we calculate the addition of the corresponding and the next local pitch values in \( V_{\text{local}} \). If the result is close to pitch average (between 95-105% of pitch average), the pitch halving has happened and therefore the relative pitch pulse is discarded from \( V_{\text{val}} \) and considered as an invalid pitch pulse in \( V_{\text{inv}} \) and then \( V_{\text{local}} \) is updated.

![Figure 3: (a) The residual speech signal including irregular pitch variations, (b) The corresponding local pitch variations.](image)

If the corresponding pitch is neither pitch doubling nor pitch halving, it is declared as an irregular pitch and the corresponding pitch-cycle is modified. Next, the pitch contour is constructed. Consider \( V_{\text{local}}(i) \) as the irregular pitch value placed between the regular pitch values \( V_{\text{local}}(m) \) and \( V_{\text{local}}(n) \), \( m < i < n \), the modified irregular pitch value is given by Eq. 14.

\[
V_{\text{local}}(i) = V_{\text{local}}(m) + \frac{V_{\text{local}}(n) - V_{\text{local}}(m)}{n - m} (i - m) \quad (14)
\]

Some exceptions may exist especially at onset and offset frames. If the irregular pitch value is the first element of the vector \( V_{\text{local}} \) at onset frames or the last element at offset frames, the
first and the last two regular pitch values before and after the irregular pitch are considered respectively for irregular pitch-value modification given in Eq. 14. Figure 4 shows the modified pitch values in comparison with the original ones for the residual speech depicted in Fig. 3(a).

Thus, at the end of this stage, the following information is provided:

- Updating the invalid pitch pulse locations in vector \( V_{\text{inv}} \). This information is used in the pitch pulse refinement described in section 3.2.
- Specifying the regular and irregular pitch values and the corresponding local pitch cycles in vector \( V_{\text{local}} \).
- Estimation of the modified pitch values instead of the irregular ones based on the smooth pitch evolutions.

3.2 Pitch Pulse Refinement

The aim of this stage is to refine high-energy invalid pitch-pulses to make more regular speech. The inputs are the residual speech signal and valid and invalid pitch pulse locations given by \( V_{\text{val}} \) and \( V_{\text{inv}} \), respectively. The proposed refinement is performed based on local-energy variations in the first regular local pitch-cycle before and after the current cycle. Therefore, firstly, the pitch cycle including the invalid pitch pulse and the previous and next regular pitch-cycles are founded by using the information of \( V_{\text{inv}} \) and \( V_{\text{val}} \). If \( n_i \in V_{\text{inv}} \) is the invalid pitch-pulse location, the cycle \( c_i \) is centred by \( n_i \) with length of \( L_i \). \( L_i \) is selected such that no other pitch pulse is included. Next, two cycles from the regular pitch cycles are searched to have maximum correlation with \( c_i \) and the same length. In order to reduce the computations, searching is performed in \([n_i-2, n_i+2]\).

Assuming \( c_m \) and \( c_n \) are the resulting cycles, we consider a moving window with length of five samples to compute the local energy for \( c_m \), \( c_n \) and \( c_i \) excluding the invalid pitch pulse energy. The assumption that refines the invalid pitch-pulse is that the local energy of the samples around the pitch-pulse (excluding the pitch-pulse) changes as a fraction of the local energy of the corresponding samples in the previous and the next cycle. Thus, we define the factor \( \beta_1 \) and \( \beta_2 \) as given by equation 15.

\[
\beta_1 = \frac{1}{M} \sum_{k=1}^{M} e_i(k) e_m(k) \\
\beta_2 = \frac{1}{M} \sum_{k=1}^{M} e_i(k) e_n(k) \\
M = \left\lfloor \frac{L_i}{5} \right\rfloor
\]

Where \( e_i(.) \), \( e_m(.) \) and \( e_n(.) \) are the local energies of the cycles \( c_i \), \( c_m \) and \( c_n \), respectively. Assuming \( E_i \), \( E_m \) and \( E_n \) as the invalid pitch-pulse energy and the local energy of the corresponding samples in \( c_m \) and \( c_n \), the irregular pitch pulse and the two samples around it are normalised by a factor \( \lambda \) such that the conditions given in equation 16 are satisfied.

\[
\frac{E_i}{E_m} = \beta_1 \\
\frac{E_i}{E_n} = \beta_2
\]
Where $\tilde{E}_{ii}$ is the energy of the modified invalid pitch-pulse:

$$\tilde{E}_{ii} = \frac{E_{ii}}{\tilde{\lambda}^2}$$

(17)

Since the conditions given in Eq. 16 may not be satisfied together using the factor $\lambda$, we define the function $\xi(.)$ as follows:

$$\xi(\lambda) = \left| \frac{E_{ii}}{E_{mi}} - \beta_1 \right| + \left| \frac{E_{mi}}{E_{ii}} - \beta_2 \right|$$

$$= \left| 1 - \frac{E_{ii}}{E_{mi}} \right| + \left| \lambda^2 \frac{E_{mi}}{E_{ii}} - \beta_2 \right|$$

(18)

The optimum normalisation factor $\lambda_{opt}$ is obtained through partial differentiating $\xi(.)$ with respect to $\lambda$:

$$\frac{\partial \xi(\lambda)}{\partial (\lambda)} = 0$$

(19)

This implies that:

$$\lambda_{opt} = \sqrt{\frac{E_{ii}}{E_{mi}} \frac{E_{mi}}{E_{ii}}}$$

(20)

As an example, two invalid pitch-pulses in the speech signal shown in figure 5(a), are refined using the described algorithm. The synthesised speech is shown in figure 5 (b).

In order to study the effect of the refinement procedure on the pitch estimation, the original and the modified speech signals are filtered using a low-pass filter and then the normalised autocorrelation is computed for pitch lags between 40-160 samples.

The results depicted in figure 6 show that the maximum of the normalised autocorrelation (which occurs at a pitch lag of 78 samples) of the modified speech is around 0.1 higher than the original one. In addition, the maximum occurred at the submultiple of the pitch (at a pitch lag of 117 samples) is more attenuated. This leads to a more reliable pitch estimate.

4. THE PROPOSED PRE-PROCESSOR EVALUATION

In order to evaluate the proposed modifier, we construct a database including words and short sentences (for both male and female speakers), where pitch evolves irregularly, from the NTT speech database [11].

The new database is used for evaluating the modifier in the following areas:

- Pitch estimation
- Voicing level estimation
- Subjective listening tests.
4.1 Effect of the Pre-processor on Pitch Estimation

The algorithm applied in the standard MELP is used to estimate the pitch values of the original and the modified speech. In [12] it is shown that more accurate estimated pitch leads to higher Pitch Prediction Gain (PPG), $\alpha$, given by:

$$\alpha = \frac{R(T)}{R(0)}$$  \hspace{2cm} (21)

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.

Due to being fractional pitch values, the PPG is computed from the up-sampled speech signals with an accuracy of 0.25 samples. Fig. 7 shows the original speech with the resulting PPGs of the modified and the original speech signals. It is observed that the PPG of the modified speech increases in comparison with the original speech. Since the frame whose pitch evolves irregularly affects the PPG of the previous and the next frames, we only compare the resulting PPG of these original frames with the relative modified frames. 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. Figure 8 indicates the results of comparison of the PPGs of the original and the modified speech for two different thresholds 0.1 and 0.2. It is observed that the proposed modifier greatly improves the PPG of the frames containing irregular pitch variations and thus the pitch can be estimated more accurately for these frames.

Figure 7: (a) Original speech and the resulting PPGs from the original and the modified speech signals, (b) The difference of the resulting PPGs.

Figure 8: The resulting PPGs of modified speech in comparison with the original one using two different thresholds, threshold = 0.1 (top) and 0.2 (bottom).
In fact, PPG shows the effect of the proposed modifier in the time domain. The modifier also provides more suitable speech spectrum for pitch estimation. Figure 9 shows the original and the modified speech frame along with their corresponding frequency spectrums. Since speech coders usually estimate the pitch from the low pass filtered speech signal with a cut-off frequency around 1KHz [13], as observed in Figure 9 (d), the modified residual spectrum is more harmonic in comparison with the original residual spectrum for frequencies between 0 to 1KHz. Thus, the pitch can be more accurately estimated from the modified speech signal. In order to show the effect of the modifier in the frequency domain the synthetic spectral matching method is used. We define the ratio of Synthetic Spectral Matching of the Modified to the Original, SSMMO, given in (22).

\[
SSMMO(n) = \frac{\tilde{E}(\tilde{\alpha}_o, n)}{E(\alpha_o, n)}
\]  

(22)

\[\tilde{E}(\tilde{\alpha}_o, n) = \sum_{m=0}^{N-1} |S_x(m, n) - \hat{S}_x(m, \tilde{\alpha}_o, n)|^2\]  

\[E(\alpha_o, n) = \sum_{m=0}^{N-1} |S_x(m, n) - \hat{S}_x(m, \alpha_o, n)|^2\]

(23)

Where \(S_x(n), \hat{S}_x(n), S_x(n)\) and \(\hat{S}_x(n)\) are the spectrums of the modified, the synthetic modified, the original and the synthetic of the \(n\)th frame speech, respectively. \(N, \omega_o\) and \(\tilde{\alpha}_o\) respectively are the FFT length, the estimated pitch frequency of the original and the modified signals. This is measured for original frames including irregular pitch variations and the corresponding modified frames. Figure 10 shows the original and the modified speech signals and the corresponding synthetic spectral matching distortions. In Figure 10-c, it is observed that for the frames including irregular pitch evolutions, the resulting distortion is significantly higher for the original compared to the modified one and consequently the SSMMO decreases (Figure 10-d). When SSMMO<1, this indicates the estimated pitch of the modified speech is more accurate than the original one and vice versa. Thus, the accuracy of the estimated pitch is measured as:

\[
\text{Accuracy of Original estimated pitch} > \text{the Modified if} \quad \text{SSMMO} < 1 - \text{TH} \\
\text{Accuracy of Modified estimated pitch} > \text{the Original if} \quad \text{SSMMO} > 1 + \text{TH} \\
\text{Same accuracy} \quad \text{if} \quad 1 - \text{TH} \leq \text{SSMMO} \leq 1 + \text{TH}
\]  

(24)

Figure 9: (a) The original LP residual, (b) the modified LP residual, (c), (d) the corresponding original and modified LP residual spectrums, respectively.

Figure 10: (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.
Accuracy of the estimated pitch of modified speech in comparison with the original one.

Threshold = 0.1

- 19% Accuracy of Modified > Original.
- 3% Accuracy of Modified = Original.
- 78% Accuracy of Original > Modified.

Accuracy of the estimated pitch of the modified speech in comparison with the original one.

Threshold = 0.2

- 35% Accuracy of Modified > Original.
- 5% Accuracy of Modified = Original.
- 60% Accuracy of Original > Modified.

Figure 11: Accuracy of the estimated pitch using the SSMMO function with two different thresholds. (Top) TH = 0.1, (Bottom) TH=0.2.

Figure 12: (a) The original speech, (b) and (c), the estimated pitch values of the original and modified male speech, respectively.

The TH is considered as a guard region to prevent small differences between the resulting distortions in our measurements. Figure 11 shows the accuracy of the estimated pitch of the modified speech in comparison with the original one using the SSMMO with TH=0.1 and TH=0.2. This shows that the modifier modifies frames containing irregular pitch evolution such that the pitch can be more accurately estimated.

Figure 13: (a) The original speech, (b) and (c), the estimated pitch values of the original and modified female speech, respectively.

In the next experiment, the effect of the modifier on the frame-to-frame pitch variations is presented. Since the modification is based on smoothing the pitch contours, it is expected that the pitch will evolve smoothly from frame-to-frame as compared to the original counterpart. Figures 12 and 13 show 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.

4.2 Effect of the Pre-processor on Voicing Level Estimation

As detailed in the last section, the modifier provides more regular speech such that the pitch can be more accurately estimated. This also affects the accuracy of the other parameters estimation, e.g., voicing level. In
the following, we show that the modifier 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.

Figures 14 and 15 show 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. Figure 16 shows the modified and the original speech spectrums in the five frequency bands along with the estimated voicing level for each band.

It is observed that in spite of being strongly voiced speech (Fig. 14-left), only the first band of the original speech is estimated as voiced, whereas the first four bands of the modified are estimated as voiced. In order to provide a quantitative measure of the accuracy of the estimated voicing level, we compute the spectral distortion given in Eq. 23, where \( \hat{S}_w(n) \) and \( \tilde{S}_w(n) \) are substituted by the original and the modified synthesised spectrums. This distortion is only computed for frames containing irregularly pitch variations and also one frame before and after these frames. Next, the SSMMO is calculated for the last four bands (for male speech only) and also for the
full band (for both male and female speech). The accuracy of the voicing decision for each band is given by Eq. 24.

Figure 17 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 and TH=0.2.

However, due to improved pitch value the spectral distortion is even lower.

4.3 Subjective Listening Tests

The database used for clean speech was used under background noise and applied to the pre-processor. An AB-test with 15 (4 trained and 11 untrained) listeners was carried out on the noise-suppressed original and the processed speech files to investigate the perceptual speech quality of the modified speech before applying to a speech coder. Results are shown in Table 1. The results obtained indicate that there is no difference in perceptual quality between the original and the modified speech under background noise for SNRs higher than 15 dB.

<table>
<thead>
<tr>
<th>Speech Type</th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean speech</td>
<td>0.2</td>
<td>11.2</td>
<td>76.3</td>
<td>12.3</td>
<td>0</td>
</tr>
<tr>
<td>25 dB SNR</td>
<td>0.2</td>
<td>10.8</td>
<td>75.9</td>
<td>12.9</td>
<td>0.2</td>
</tr>
<tr>
<td>20 dB SNR</td>
<td>0.1</td>
<td>10.2</td>
<td>74.6</td>
<td>14.7</td>
<td>0.4</td>
</tr>
<tr>
<td>15 dB SNR</td>
<td>0.1</td>
<td>9.3</td>
<td>71.6</td>
<td>17.8</td>
<td>1.2</td>
</tr>
<tr>
<td>10 dB SNR</td>
<td>0</td>
<td>7.8</td>
<td>66.4</td>
<td>23.4</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 1: Modified speech vs. original speech

<table>
<thead>
<tr>
<th>Speech Type</th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean speech</td>
<td>4.2</td>
<td>51.4</td>
<td>30.1</td>
<td>12.2</td>
<td>2.1</td>
</tr>
<tr>
<td>25 dB SNR</td>
<td>3.4</td>
<td>50.2</td>
<td>31.2</td>
<td>12.9</td>
<td>2.3</td>
</tr>
<tr>
<td>20 dB SNR</td>
<td>2.8</td>
<td>47.3</td>
<td>32.6</td>
<td>14.2</td>
<td>3.1</td>
</tr>
<tr>
<td>15 dB SNR</td>
<td>2.2</td>
<td>43.1</td>
<td>33.6</td>
<td>17.3</td>
<td>3.8</td>
</tr>
<tr>
<td>10 dB SNR</td>
<td>0.8</td>
<td>30.3</td>
<td>36.2</td>
<td>25.9</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Table 2: Modified speech + MELP vs. MELP alone

In the next experiment, the noisy speech signals were applied as the inputs to standard MELP 2.4 kbps and an A vs. B comparison test was carried out on the synthesised speech files. The results are shown in Table 2. This shows that the pre-processor in combination with the MELP provides...
significantly better perceptual speech quality than the standard MELP at SNRs higher than 15 dB.

4. CONCLUSIONS

In this paper the current pre-processors and techniques for pitch-estimation improvement were addressed. We proposed a new pre-processor, which modifies the residual signal to make more regular speech. The modification is performed on the cycles containing irregular pitch variation where these cycles are marked using a smooth pitch contour. This procedure is based on inserting or discarding a segment. The optimum segment is searched based on increasing the correlation between the modified and the previous cycle and minimisation of the resulting discontinuities at connection points. The pitch prediction gain and spectral distortion were used as quantitative measures to demonstrate that the benefits of applying the pre-processor are to reduce the inaccuracy of the pitch and voicing level estimation in speech coders. The subjective listening tests show that the pre-processor maintains the original speech quality and can therefore be used in combination with any speech coder. In addition, it was shown that the proposed pre-processor in combination with the standard MELP 2.4Kbps provides significantly better quality than MELP alone. The performance of the pre-processor under background noise was tested. Vehicle noise with SNRs of 10 to 25 dB was considered as the background noise. Results of these experiments show the robustness of the pre-processor at SNRs higher than 15 dB.

REFERENCES