Adaptive Long-Term Predictive Analysis of Disordered Speech

A. Kacha (1), F. Grenez (1), F. Bettens (1), J. Schoentgen (2)

(1) Department Waves and Signals, Université Libre de Bruxelles, Brussels, Belgium
(2) Laboratory of Experimental Phonetics, Université Libre de Bruxelles, Brussels, Belgium

Abstract

A time-varying long-term bi-directional linear predictive analysis is proposed in the context of the acoustic assessment of disordered speech. It is shown that performing a forward and backward long-term linear prediction of each speech sample and retaining the minimal overall prediction error as a cue of vocal dysperiodicity obtains a signal-to-noise ratio that correlates with the perceived degree of hoarseness. The coefficients of the time-varying long-term linear predictive model are estimated adaptively by means of a recursive least squares algorithm.

1. Introduction

Acoustic analyses of speech signals are popular in the framework of the clinical evaluation of voice because the analysis is non-invasive and documents quantitatively the degree of hoarseness perceived by the clinician. One acoustic marker of hoarseness is the so-called signal-to-noise ratio (SNR) [1], [2]. In the context of the assessment of disordered speech, noise refers to dysperiodicities that are detected in the speech waveform, including additive noise owing to turbulence and modulation noise owing to perturbations of the glottal excitation signal.

Several techniques have been proposed to estimate the signal-to-noise ratio in disordered speech [3], [4]. One technique is conventional linear predictive modeling [5] referred to as short-term linear prediction. In short-term linear prediction, the current speech sample value is estimated as a linear combination of past sample values. It has been applied to the assessment of the speech of dysphonic speakers in combination with a long-term predictive analysis [6]. The basic idea of long-term prediction which was introduced originally in the context of speech coding [7] is to remove the distant waveform redundancies by estimating the current sample value on the basis of the previous far-sample values. It has been found that combining both short-term and long-term linear predictive models results in a smaller overall prediction error.

Long-term prediction per se (in combination with a short-term linear prediction stage or not) is an attractive analysis scheme for disordered speech. Indeed, when the speech signal is cyclic and the cycle amplitudes change smoothly, it is possible to predict the present cycle on the basis of some distant previous cycles. The prediction distance is chosen such that it minimizes the mean square long-term prediction error. The prediction distance is identical to the length of the glottal cycle or a multiple thereof in the case of voiced and pseudo-periodic speech sounds. For unvoiced speech segments, the prediction distance cannot be interpreted in terms of the glottal cycle length. It remains meaningful for computational purposes, however.

If the signal is periodic, the present cycle can be perfectly predicted by means of one of the previous cycles. When the amount of vocal dysperiodicity increases, the long-term prediction error increases. The long-term prediction error is therefore a cue of the amount of vocal dysperiodicity and a signal-to-noise ratio can be defined involving the original signal and the long-term linear prediction error. Combining short-term and long-term predictive models does not, however, always yield SNR values that correlate perfectly with the perceived degree of hoarseness or measured speech dysperiodicity. A reason is that the short-term linear prediction is segment-dependent and speaker-dependent, i.e., speaker or speech properties that are not related to any dysperiodicities due to voice disorders may increase the prediction error [8]. Moreover, it has been observed that the long-term prediction error is inflated at the beginning of the recording interval and at the transition between voiced and unvoiced segments.

To overcome these problems, long-term bi-directional prediction has been proposed [8], [9]. In this approach, only the long-term predictive model is used, the analysis being performed in the backward and forward directions. The backward prediction was carried out by time-reversing the signal and therefore performing a forward linear prediction on the reversed signal, which assumes that the entire signal is available.

Although the results obtained by long-term bi-directional prediction appear to be promising, the assumption that the signal is locally stationary may be a limitation in the framework of the analysis of continuous speech, especially while attempting modeling transitions that are shorter than the analysis window.

The aim of this presentation is therefore to extend the long-term bi-directional prediction to the general case of time-varying signals, omitting the assumption of local stationarity. The time-reversing of the signal is avoided by using a parallel prediction scheme that combines forward and backward prediction filters like in two-sided linear prediction [11] but in a competitive manner which means that at each time instant, only a limited future segment of the signal has to be available. The coefficients of the time-varying bi-directional linear predictive model are estimated adaptively for each signal sample by using the recursive least squares (RLS) algorithm. The article is organized as follows. In section 2, the time-varying long-term prediction approach is presented. Since, the
involved time-varying linear predictive model differs from the conventional one, the RLS algorithm is revised accordingly. The corpus used in the experimental evaluation of the proposed approach is described. In section 3, the results are presented. The amount of estimated dysperiodicities is summarized for each signal by means of an acoustic marker, which involves the long-term bi-directional prediction error as well as the clean signal. Finally, in section 4, perspectives are discussed.

2. Analysis

To remove the far-sample (i.e. inter-cycle) redundancies and obtain an estimate of the cycle-to-cycle dysperiodicity [7], a long-term linear prediction has been proposed. In the framework of speech coding this long-term analysis is applied to the prediction error of the conventional short-term analysis. In this article, the long-term analysis is applied to the speech signal directly. Due to the cycle-to-cycle prediction, the long-term predictive model indeed offers a tool for isolating differences between neighboring cycles. The long-term prediction error thus forms a cue that is expected to correlate with the perceived degree of hoarseness of disordered speech.

2.1. Time-invariant long-term linear prediction

The purpose of the long-term linear prediction is to remove the far-sample redundancies of the signal by acting on distant-sample waveform similarities [6], [7], [8] and [9]. Let \( x(n) \) be a stationary discrete-time zero-mean signal. The time-invariant long-term linear predictor of \( x(n) \) is given by

\[
\hat{x}(n) = -\sum_{i=0}^{m+1} a_i x(n-P-i)
\]

where \( m+1 \) is the order of the model; \( a_i, i = 0, \ldots, m \), are the parameters to be computed; and \( P \) is the prediction distance that is related to the vocal cycle length in the case of voiced speech sounds. The resulting long-term prediction error is as follows:

\[
e(n) = x(n) + \sum_{i=0}^{m} a_i x(n-P-i)
\]

Order \((m+1)\) is typically equal to 3. The motivation for involving more than one speech sample in the prediction is that the actual cycle length does not necessarily agree with an integer number of sampling steps. If the signal is non-stationary, it is divided into frames of short duration (e.g. 2.5 ms) and each frame is modeled by (1). The coefficients of the model are calculated by minimizing the following squared error.

\[
\varepsilon^2 = \sum_{n=-\infty}^{\infty} e^2(n)
\]

Error \( e(n) \) is obtained by windowing the long-term error by some window. As for the conventional short-term linear prediction, depending on the windowing, the prediction error minimization is performed by means of the autocorrelation method or the covariance method. Also, a modified Burg method has been proposed as an alternative to ensure the stability of the synthesis filter in the context of speech coding [7].

The choice of the optimum value of the lag \( P \) was addressed in [7]. It is possible to perform an exhaustive search for the optimal lag that minimizes the prediction error. A computationally less expensive method has been proposed in [8], [9]. It consists in assigning prediction distance \( P \) to the lag for which the normalized cross-correlation function between the current and a delayed signal frame is a maximum. The delay has been varied from 2.5 ms to 20 ms, i.e. likely values of the glottal cycle length.

2.2. Time-varying long-term linear prediction

The time-varying long-term linear predictive model is derived from (1) by assuming that the coefficients of the model and the lag \( P \) are time-dependent

\[
\hat{x}(n) = -\sum_{i=0}^{m} a_i(n)x(n-P(n)-i)
\]

The corresponding long-term prediction error becomes

\[
e(n) = x(n) + \sum_{i=0}^{m} a_i(n)x(n-P(n)-i)
\]

A property of this model is that the signal has not to be assumed to be stationary over the short time interval that is equal to the duration of the analyzing frame. The model is therefore suited for the analysis of continuous speech and for modeling transitions between different phonetic segments. The time-varying long-term prediction coefficients are updated for each time sample by means of a revised RLS algorithm. The minimization of the weighted cumulative squared error [10],

\[
J(n) = \sum_{k=0}^{n} \lambda^{-k} e^2(k)
\]

where \( \lambda \) is a forgetting factor, yields the set of equations that generate iteratively the time-varying coefficients. In matrix form, the parameter estimates at time \( n \) are estimated from those at time \( n-1 \) as follows.

\[
A(n) = A(n-1) - R^{-1}(n) X(n) e(n)
\]

where

\[
A(n) = [a_0(n), \ldots, a_m(n)]^T
\]

\[
X(n) = [x(n-P(n)), x(n-P(n)-1), \ldots, x(n-P(n)-m)]^T
\]

\[
R(n) = \sum_{k=0}^{n} \lambda^{-k} X(k) X^T(k)
\]

and \( T \) denotes the transposition operator. The parameter \( \lambda \) must be close but less than 1 to account for the nonstationarity of the signal. In this paper, it has been fixed to 0.97. The inverse of expression \( R(n) \) is updated at each time instant on the base of the matrix inversion lemma. In this article, the optimal prediction distance \( P(n) \) is determined via the short-term cross correlation function at time instant \( n \) between the signal and a delayed signal frame of a duration of 2.5 ms. The prediction distance is assigned to the delay for which the cross-correlation function is maximal. The delay has been comprised between 2.5 ms and 20 ms.
2.3. Bi-directional time-varying long-term linear prediction

Since the purpose of linear prediction-based analysis of disordered speech is to isolate vocal dysperiodicities by minimizing a cumulative squared error, both forward and backward time-varying linear predictions may be performed and the recently best predictor of competitive predictors may be kept at each time instant as the optimal predictor in the sense of the minimal absolute value of the prediction errors. In practice, the purpose of the bi-directional analysis is to remove clinically spurious prediction errors that occur at the beginning of the recording interval and at the boundaries between the phonetic segments in connected speech. It is indeed not possible to predict distant speech samples across phonetic boundaries, because the cycle shape depends on the phonetic identity of each segment. Bi-directional analysis entails that speech samples are either predicted or retrodicted depending on which direction gives rise to the smallest prediction error. Prediction across segment boundaries is thus avoided and the observed error is likely to be caused by vocal perturbations rather than by the evolving phonetic identity of the speech segments.

The backward time-varying linear predictor uses the future samples to estimate the present sample. It can be expressed as

\[ \hat{x}(n) = - \sum_{i=0}^{m} b_i(n)x(n + P_b(n) + i) \]  

where \( b_i \) and \( P_b \) are the coefficients and the distance corresponding to the backward time-varying linear predictor, respectively. The processing performed for model (4) applies equally to model (11). The backward time-varying prediction error is given by

\[ e_b(n) = x(n) + \sum_{i=0}^{m} b_i(n)x(n + P_b(n) + i) \]  

The coefficients \( b_i(n) \) are calculated via Eqs (7)-(10) that have been revised according to the previous modifications of the prediction error \( e_b(n) \) and the vector of data samples. At each time instant, the estimation is performed with a delay of \( (P_{max}+m) \), i.e., to estimate the prediction error at time instant \( n \), samples up to time instant \( (n + P_{max}+m) \) are needed.

2.4. Estimation of SNR

The acoustic marker is the so-called signal-to-noise ratio that summarizes the amount of dysperiodicities for each signal as follows.

\[ \text{SNR}_{DB} = 10 \log \frac{\sum_{n=0}^{N-1} \hat{x}^2(n)}{\sum_{n=0}^{N-1} e_m^2(n)} \]  

where \( e_m(n) \) is either \( e(n) \) or \( e_b(n) \) depending on which has the minimum absolute value and

\[ \hat{x}(n) = x(n) - e_m(n) \]  

with \( N \) denoting the number of samples.

2.5. Corpus

The first corpus comprises 1-second stationary fragments of vowel [a] sustained by 38 healthy speakers (22 male and 16 female) and 51 dysphonic speakers (19 male and 32 female). The sampling frequency has been 20 kHz. The degree of hoarseness has been determined by five judges on the base of a visual inspection of the spectrograms. Each judge has assigned a degree between 0 and 4. Consequently, the overall degrees of perceived hoarseness have been comprised between 0 and 20. Spectrograms that have been considered to be noisy or clean have thus been assigned high or low degrees of hoarseness respectively.

The second corpus is a subset of the signals corresponding to the French sentence “il est sorti avant le jour” uttered by a female speaker [12]. The sampling frequency has been 44.1 kHz. The signals are labeled in an increasing order of hoarseness as “modal”, “rough 1”, “rough 2”, “rough 3” and “whisper”.

3. Results

Figure 1 (a) shows a fragment of a vowel [a] that has been assigned an overall degree of hoarseness of 19. Figure 1(b) shows the corresponding time-varying long-term bi-directional prediction error. The SNR value for that signal has been equal to 8.2 dB, which is an indication of a large amount of vocal dysperiodicity.

The SNR value of each vowel has been calculated and plotted versus the corresponding degree of hoarseness as shown in Figure 2. The scattergram suggests that when the degree of perceived hoarseness increases/decreases, the corresponding SNR marker value decreases/increases. This is confirmed by Spearman’s rank correlation coefficient, which is equal to -0.8.

Figure 3 (a) displays the phonetic segment [il] corresponding to the French sentence of the second corpus, labeled “modal” and figure 3 (b) shows the corresponding long-term bi-directional prediction error. The SNR values of the different signals are given in Table 1.

4. Discussion and conclusion

An adaptive time-varying long-term bi-directional predictive model has been proposed for the analysis of disordered speech and tested on vowel fragments and continuous speech. The experimental results indicate that the formulation of the problem of dysperiodicities estimation as a long-term prediction problem allows to capture the voice quality of pathological speech from sustained vowels as well as from continuous speech. By using an adaptive model, the assumption of local stationarity can be omitted which means that the segmentation of the signal into short frames is avoided and the transitions between different phonetic segments are well modeled. It has been found that the calculated acoustic marker values are highly correlated with a classification according to the perceived degree of hoarseness. Table 1 shows that the estimated SNRs are in good agreement with the quality of voice. Indeed, the signal labeled modal which corresponds to a normal speaker is characterized by a high SNR and that labeled whisper correspond to a highly dysphonic speaker is characterized by a small SNR. It is worth noting that the distance of prediction was assumed to be a multiple of the sampling period. The results may be improved by using an interpolation procedure to find a more accurate
value of the prediction distance. Possible applications of the proposed approach are therefore the real-time tracking of evolving vocal dysperiodicity marker values in connected speech.

Figure 1: Speech signal (a) whose degree of hoarseness equals 1 and its long-term bi-directional prediction error (b).

Figure 2: SNR value versus perceived degree of hoarseness.

Table 1: SNR values of the continuous speech corpus.

<table>
<thead>
<tr>
<th>Signal</th>
<th>SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modal</td>
<td>25.3</td>
</tr>
<tr>
<td>Rough1</td>
<td>17</td>
</tr>
<tr>
<td>Rough2</td>
<td>13.5</td>
</tr>
<tr>
<td>Rough3</td>
<td>10.1</td>
</tr>
<tr>
<td>Whisper</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Figure 3: Fragment of a modal speech signal (a) and its long-term bi-directional prediction error (b).

5. References