

Estimation of Vocal Noise in Running Speech by means of Bi-directional Double Linear Prediction

F. Bettens(*), *F. Grenez* (*), *J. Schoentgen* (**)

Service Electricité Générale (*), Laboratory of Experimental Phonetics (**)
Université Libre de Bruxelles, Belgium
(**) National Fund for Scientific Research, Belgium
jschoent@ulb.ac.be

Abstract

The presentation concerns forward and backward double linear prediction of speech with a view to the characterization of vocal noise due to voice disorders. Bi-directional double linear prediction consists in a conventional short-term prediction followed by a distal inter-cycle prediction that enables removing inter-cycle correlations owing to voicing. The long-term prediction is performed forward and backward. The minimum of the forward and backward prediction error is a cue of vocal noise. The minimum backward and forward prediction error has been calculated for corpora involving connected speech and sustained vowels. Comparisons have been performed between the estimated vocal noise and the perceived hoarseness in steady vowel fragments, as well as between the estimated vocal noise in connected speech and sustained vowels produced by the same speakers.

1. Introduction

Voice disorders refer to abnormal conditions of the glottal excitation signal that is generated by the vibrating vocal folds and the pulsatile glottal airflow. Voice disorders may be the product of disease or accidents that affect the larynx. One frequent symptom of voice disorders is an increase of the dysperiodicity of voiced speech sounds.

The observation that abnormal dysperiodicity is common in voice disorders has led to the routine use of speech analysis software in ENT-departments with a view to the calculation of numerical cues that summarize the amount of dysperiodicity in voiced speech sounds. The most often used speech segments are stationary fragments of sustained vowels, namely [a] [1]. The reason for the almost exclusive focus on stationary speech fragments is that extant analysis heuristics involve the detection of the lengths and amplitudes of individual speech cycles. This detection fails occasionally on speech fragments that are not pseudo-periodic and pseudo-stationary.

When owing to severe disorders the vowel fragments are not (pseudo)steady or (pseudo)periodic, conventional clinical analysis tools may output cue values that are not reproducible or difficult to interpret. In addition, clinicians have argued that analyses of sustained speech sounds that include onsets and offsets, as well as analyses of running speech fragments may reveal voice disorders more readily than stationary fragments [2,3]. The reasons are that in connected speech the vibration of the vocal folds must be switched on and off, the larynx may be raised and lowered, and the vibration of the

folds must be maintained in the presence of variable acoustic loads.

Clinical speech analysis tools that enable unsteady or aperiodic speech fragments to be analyzed are uncommon, however [4,5]. Reference [4] concerns the use of electroglottograms with a view to clinical analyses of running speech. Reference [5] is about the use of short-term spectra with a view to the description of dysperiodic read speech. The analysis is based on the segregation of the pseudo-harmonics and inter-harmonics in short-term spectra. The dysperiodicity cues are not always accurate, however, because of possible erroneous detections or omissions of spectral pseudo-harmonics. Also, spectral dysperiodicity cues are unduly influenced by the vocal tract transfer function, as well as by frequency modulation noise [6,7].

One recent proposal enables analyzing vocal noise in voiced speech that is not pseudo-periodic [8]. The analysis scheme consists of two steps [9]. In a first step, the short-term correlations in the speech signal, which are due to the transfer function of the vocal filter, are removed by means of conventional linear predictive analysis. In a second step, any inter-cycle correlations that remain in the short-term prediction error owing to voicing are modeled and removed by means of a distal linear prediction filter.

This double linear prediction relies on the assumption that when the speech signal is cyclic and when the cycle amplitudes change smoothly, the present cycle can be predicted from the previous one. This prediction is perfect (that is, the prediction error is zero) when the speech signal is periodic. The inter-cycle prediction error increases when the speech signal is pseudo-periodic. The prediction error increases still more when the speech cycles evolve randomly, and it is identical with the signal when the signal is de-correlated noise.

Desirable properties of the distal prediction are that the prediction error evolves smoothly with the degree of vocal dysperiodicity, and a small increase in the vocal noise is guaranteed to produce a small increase in the prediction error, because of the linearity of the analysis. Actually, a jumpy increase of the measured perturbations owing to erroneous omissions or detections of cycle markers cannot occur in the context of distal linear prediction, because it is not based on the explicit knowledge of the glottal cycle lengths. Distal linear prediction instead relies on the knowledge of the distal prediction error, which can be computed for voiced, unvoiced and transient speech sounds. The distal prediction is indeed performed over a distance that minimizes the prediction error. This distance necessarily agrees with the length of a glottal cycle or a multiple thereof when the speech sound is voiced

and periodic. When the sound is not voiced, the distance that minimizes the distal prediction error remains meaningful for computational purposes, but is not interpreted in terms of the glottal cycle length.

Although double linear prediction has been used to analyze read speech in a clinical framework [8], it can be shown that the distal prediction error may locally be an artifactual cue of vocal noise when the speech stream comprises both voiced and unvoiced speech sounds. The reason is that the distal prediction is unidirectional, which means that sound fragments that are to the left are used to predict sound fragments that are to the right. In the vicinity of phonetic boundaries or other borders, the distal prediction error may therefore be abnormally large. This is because it is not possible to predict sound fragments at the beginning of a recording interval, or across the boundary between two speech sounds that are phonetically distinct.

The purpose of the present article is to show how distal prediction may be augmented so as to turn the distal prediction error into an artifact-free cue of vocal noise, while keeping the desirable properties that have been outlined above, i.e. the linearity, the ability to deal with periodic as well as aperiodic speech sounds, and the leaving out of an explicit estimation of the vocal frequency of voiced sounds.

2. Analysis

A solution to the problem that has been outlined in the previous section may be based on the following observation. Typically, a speech sound consists of an onset followed by a steady fragment followed by an offset. This means that in a left-right prediction scheme, the onsets are expected to predict the middle fragments, which are expected to predict the offsets. But the offsets are not expected to predict the onsets of the following speech sounds, which are phonetically distinct. Conversely, in a right-left retrodiction scheme, the offsets are expected to retrodict the middle fragments, which are expected to retrodict the onsets. But, the onsets are not expected to retrodict the offsets of the preceding sounds across their common boundary. This would suggest that to the right and left of a speech sound boundary, the vocal noise is best represented by the distal retrodiction and prediction errors respectively.

We therefore propose to use as a cue of vocal noise, in place of the forward prediction error, the minimum of the backward and forward prediction errors. The forward and backward prediction errors are compared analysis window by analysis window and the prediction error with the least energy is assigned to the vocal noise. In [8] and [9], the proximal intra-cycle prediction is called the short-term prediction and the distal inter-cycle prediction is called the long-term prediction. This terminology is used hereafter. Also, the present analysis keeps the short-term linear prediction, although it is not mandatory in the context of bi-directional long-term linear prediction.

2.1. Short-term linear prediction

The purpose of this stage was to obtain the short-term prediction error. Short-term linear prediction is the conventional linear prediction. Discrete quantities $x(n)$ and $e_p(n)$ are the speech and residue signals respectively, symbol N is the order and numbers a are the coefficients of the linear prediction filter.

$$e_p(n) = x(n) + \sum_{i=1}^N a_i x(n-i) \quad (1)$$

The analysis was performed without overlap by means of a sliding rectangular window of 10 ms. The filter coefficients were computed by means of Burg's method [10]. The filter order was 24, taking into account a sampling frequency of 20 kHz.

2.2. Long-term linear prediction

The purpose of this stage was to obtain the long-term prediction error, which was free of any cyclic correlation [9]. Model (2) is referred to as double linear prediction because it formally involves a short-term linear prediction of the speech signal, which enables short-term correlations to be removed, and a long-term linear prediction of the short-term prediction error (1), which enables inter-cycle correlations to be removed.

Long-term linear prediction is the prediction of a future sample at a distance P from the present sample. Signal $e_d(n)$ is the long-term prediction error.

$$e_d(n) = e_p(n) + \sum_{i=0}^{Np} b_i e_p(n-P-i) \quad (2)$$

According to [9], coefficients b were computed by the Burg method after model (2) had been reformulated in the framework of a lattice structure. The calculation was carried out without overlap by means of a rectangular window the length of which was 2,5 ms. Parameter Np was equal to 2 [8].

For a given analysis window position, distance P was fixed as follows. The correlation between the content of the present window and the content of a shifted window was computed with the inter-window lag varying between 2,5 ms at least and 20 ms at most. For a given lag, the correlation coefficient was obtained via the scalar product of the lagged and un-lagged windows, and a normalization by the windowed signal energy, so as to compensate for drifts in the signal amplitude. Distance P was assigned to the lag for which the inter-window correlation was a maximum.

2.3. Bi-directional double linear prediction

We have argued above that phonetic boundary artifacts can be removed by performing a backward and forward linear prediction, and by assigning the minimum of the backward and forward prediction errors to the vocal noise.

In practice, the forward short-term and long-term linear predictions were performed as outlined above [8,9]. The outcome of the forward double linear prediction was the forward decorrelated residue with large spurious error transients at the beginning of the recording, and at the beginning of voiced speech sounds that were preceded by unvoiced ones (and vice versa). This forward double prediction error was stored for further processing.

The short-term residue was then time-reversed, that is, the last sample of the recording interval became the first and vice versa. After that, a second long-term linear prediction was performed, which was a backward prediction because of the time reversal. Care was taken to position the center of the mobile analysis window at the same positions as during the forward prediction. A second time reversal gave the backward decorrelated residue with large error transients at the end of

the recording and near the offsets of voiced sounds that were followed by unvoiced ones (and conversely). This backward double prediction error was stored for further processing.

The last stage consisted in comparing the forward and backward decorrelated residuals, analysis window by analysis window, and assigning the prediction error with the least energy to the vocal noise. Taking the minimum removed artifacts because during forward prediction the offsets of phonetic segments were correctly predicted, whereas during backward linear prediction the onsets were correctly predicted.

The length of most artifactual transients was typically one glottal cycle. The removal of artifactual transients was therefore successful for any phonetic segment whose length was equal to two glottal cycles at least, a condition that is met by a great majority of speech sounds in running speech.

2.4. Excitation to noise ratio

A predicted excitation to prediction noise ratio, ENR , was computed as follows. Symbol M is the number of samples comprised within a recording interval. The numerator of the ratio is the energy of the predicted short-term residue (i.e. the “clean” excitation) and the denominator is the energy of the least bi-directional double prediction error (i.e. the vocal noise). The value of the ratio is therefore comprised within the interval $(-\infty, +\infty)$, with large positive values indicating feeble vocal noise.

$$ENR = 10 \log \frac{\sum_{n=1}^M [e_p(n) - e_{d,l}(n)]^2}{\sum_{n=1}^M e_{d,l}^2(n)} \quad (3)$$

3. Corpora

A first corpus comprised 1-second long steady portions of the vowel [a] sustained by 89 normophonic or dysphonic speakers. The degree of visual hoarseness was established by a jury of five experts on the base of a 5-point hoarseness scale established by Yanagihara [11]. The perceived hoarseness therefore varied between 0 (clean speech) and 20 (very hoarse speech) [6]. A second corpus comprised logatom sequences [apa] and vowels [a] (including onsets and offsets) produced by 20 normophonic or dysphonic speakers [12]. The sampling frequencies for the two corpora were 20 kHz and 24 kHz respectively.

4. Results

Figure (1) displays a sustained vowel fragment together with the short-term prediction error, least double linear prediction error, and forward double linear prediction error (in that order from the top). The latter shows a large initial error transient owing to the inability to predict initial speech fragments from the outside of the recording interval. Similar spurious error transient may be observed in the vicinity of phonetic boundaries owing to the inability to predict voiced speech fragments from unvoiced ones (and vice versa).

Figure (2) displays the excitation to noise ratio (on the vertical axis) and the corresponding degree of perceived hoarseness (on the horizontal axis) for steady 1-second long vowel fragments sustained by 89 normophonic and dysphonic

speakers. The purpose of the perceptual classification was to inform independently from the calculated noise ratio (3) about the severity of the dysperiodicity of each vowel fragment. The value of the coefficient of correlation between excitation to noise ratio and degree of perceived hoarseness was -0,82.

Finally, Figure (3) displays the values of the excitation to noise ratios for logatom sequences [apa] on the vertical axis and sustained vowels [a] (including onsets and offsets) on the horizontal axis. Logatoms [apa] and vowels [a] were produced by the same set of 20 normophonic and dysphonic speakers. The value of the coefficient of correlation between the excitation to noise ratios for [apa] and [a] was 0,93. The straight line is the Figure’s bisector.

The perceptual evaluation of the [apa] sequences was too difficult, because they were too short [12]. One may however argue that the vocal noise in different speech fragments produced by the same speaker must be correlated. The observation of a high correlation between [a] and [apa] therefore suggests that the removal of the artifacts owing to the voiced/unvoiced transitions in [apa] was successful. Otherwise the scatter of the data would have been higher or the regression line would have been far from the bisector of Figure (3).

The observation of smaller values for the excitation to noise ratio in Figure (3) than in Figure (2) is expected, because the data in Figure (3) involve the vowel onsets and offsets, which decrease the excitation to noise ratio.

5. Discussion and conclusion

Bi-directional double linear prediction has the advantage over forward double linear prediction that any artifactual inter-segment transients are removed in the least double linear prediction error. This is a benefit for clinical applications when the least double linear prediction error is used as a cue of vocal noise. The inability to obtain the latter in real-time is not a disadvantage in a clinical framework, because a short delay between the recording and the output of the analysis result is acceptable.

Also, in the context of forward double linear prediction, short-term linear prediction is essential. This is because otherwise any phonetic segment boundary would give rise to spurious transients in the forward double prediction error. The reason is that the timbre of the present speech sound does not predict the timbre of the following one. Short-term linear prediction removes the timbre-related short-term correlation, however. Spurious transients in the forward double prediction error are therefore expected only where voiced speech sounds border on unvoiced ones.

In the context of bi-directional linear prediction, the short-term linear prediction is an option, but not an obligation. The reason is that in bi-directional double linear prediction analysis, speech fragments must not be predicted across segment boundaries. This means that the short-term prediction can be dropped and the bi-directional long-term prediction can be performed alone. Bi-directional long-term linear prediction analysis does not give rise to spurious inter-segment transients in the least long-term prediction error, because a segment’s onset and offset may be respectively retrodicted and predicted from its middle fragment. Bi-directional long-term linear prediction of speech (in place of the short-term prediction error) is attractive because it mimics

numerically the visual back and forth scanning performed by a human expert who inspects graphs of speech signals.

6. References

- [1] Kent R. D., Ball M. J (Eds), *Voice Quality Measurement*, Singular, Thomson Learning, San Diego, 2000.
- [2] De Krom G., Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments, *J. Speech, Hearing Res.*, 38, 4, 1995, 794-811.
- [3] Parsa V., Jamieson D. G., Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech, *J. Speech, Lang., Hear. Res.*, 44, 2001, 327-339.
- [4] Abberton E., Fourcin A., Electrolaryngography, in *Instrumental Clinical Phonetics*, Ball M. J. & Code C. (Eds), Whurr Publishers, London, 1997, 119-148.
- [5] Klingholtz F., Acoustic Recognition of voice disorders: A comparative study of running speech versus sustained vowels, *J. Acoust. Soc. Am.*, 87, 5, 1990, 2218-2224.
- [6] Schoentgen J., Bensaid M., Bucella F., Multivariate statistical analysis of flat vowel spectra with a view to characterizing dysphonic voices, *J. Speech, Lang., Hear. Res.*, 43, 2000, 1493-1508.
- [7] Schoentgen J., Spectral Models of additive and modulation noise in speech and phonatory excitation signals, *J. Acoust. Soc. Am.*, 113, 1, 2003, 553-562.
- [8] Qi Y., Hillman R. E., Milstein C., The estimation of the signal to noise ratio in continuous speech in disordered voices, *J. Acoust. Soc. Am.*, 105, 4, 1999, 2532-2535
- [9] Ramachandran R. P., Kabal P., Pitch prediction filters in speech coding, *IEEE Trans. Acoust., Speech, Sig. Proc.*, 37, 1989, 467-478.
- [10] Burg J. P., A new analysis technique for time series data, in *Modern Spectrum Analysis*, Childers D. G. (Ed.), IEEE Press, NY, 1978, 42-48.
- [11] Yanagihara N., Significance of harmonic changes and noise components in hoarseness, *J. Speech, Hearing Res.*, 10, 1967, 531-541.
- [12] Clérin C., *Utilisation de la voyelle soutenue et de la parole continue pour l'évaluation vocale : comparaison à partir d'un échantillon de voix dysphoniques*, unpublished Master thesis, Univ. Libre de Bruxelles & Univ. Catholique de Louvain, 2001.

Figure 1 (from the top) : Steady vowel fragment, short-term prediction error, least bi-directional double prediction error (i.e. vocal noise) and forward double prediction error.

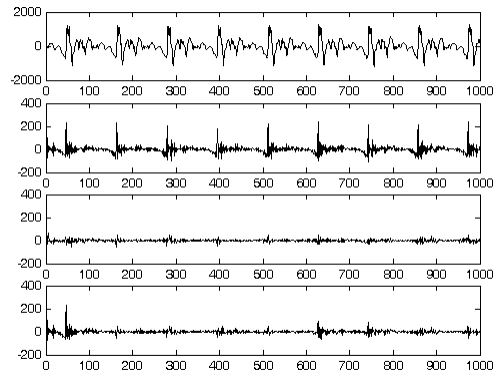


Figure 2 : Excitation to noise ratio (vertical axis) and degree of perceived hoarseness (horizontal axis); steady vowel fragments [a].

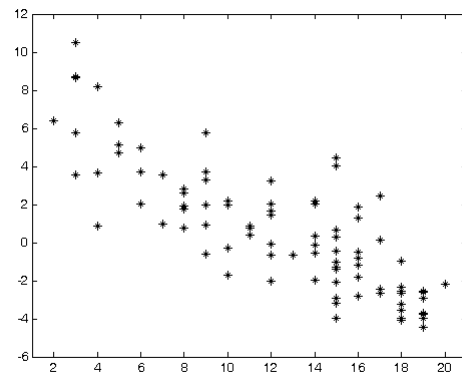


Figure 3 : Excitation to noise ratio for logatoms [apa] (vertical axis) and vowels [a] (horizontal axis).

