Dysphonic Speech Analysis Using Generalized Variogram

A. Kacha [1], F. Grenez [1], J. Schoentgen [2], K. Benmahammed [3]

[1] Department Waves and Signals, Université Libre de Bruxelles, Brussels, Belgium
[2] Laboratory of Experimental Phonetics, Université Libre de Bruxelles, Brussels, Belgium ; National Fund for Scientific Research, Belgium
[3] Laboratory of Intelligent systems, Université de Setif, Algeria
E-mail: akacha@ulb.ac.be

ABSTRACT

Acoustic analyses of speech signals are popular in the framework of the clinical evaluation of voice and diagnose of disease. We propose a new strategy for dysphonic speech analysis that extracts vocal dysperiocities by using a generalized form of the variogram. The generalized variogram allows to overcome the inherent drawbacks of both long-term and short-term linear prediction formulations widely used in disordered speech analysis. The proposed approach uses a forgetting factor to account for the nonstationarity nature of the speech signal. Experimental results show that the proposed approach outperforms the double prediction-based technique.

1. INTRODUCTION

Acoustic analyses of speech signals are popular in the framework of the clinical evaluation of voice because the analysis is noninvasive 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]. Several diseases result in a decreasing energy of the harmonic structure of the spectrum in detriment of that of the nonharmonic structure. 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 caused by a malfunction of the vocal folds [2, 3]. Speech signal may contain an unvaluable information on the degree of hoarseness. Although there are various medical conditions that can affect the voice, most of the disorders originate from the vocal system and frequently result in an increase in the dysperiodicity of voiced speech sounds providing a motivation to quantify the amount of this dysperiodicity by estimating the SNR.

Most approaches for dysphonic speech analysis are based on the periodicity of the vocal folds vibration [4, 5]. Even these methods have been successfully applied to sustained vowels, they exhibit a lack of robustness and accuracy when they are applied to the estimation of vocal noise in continuous speech or vowels including onsets and offsets. Indeed, these techniques require a stationary portion of the speech signal either for the mathematical model to be valid or for an accurate measurement of the acoustic parameters of the analyzed speech signal. Up to date, there is comparatively a small number of studies conducted on continuous speech [6, 7, 8]. This is due primarily to the difficulty in continuous dysphonic speech to isolate the individual speech cycles and the individual harmonics which gives rise to a biased acoustic marker of vocal noise. In [7], the authors attempted to avoid this drawback by using a double prediction-based method where two analysis stages have been used in cascade. The first stage is composed of a conventional linear prediction modeling refereed to as short-term linear prediction. The aim of the linear prediction modeling is to remove the near-sample waveform redundancies by estimating the current speech sample value as a linear combination of the past values. The second stage performs a long-term predictive modeling on the signal residue obtained at the output of the first stage.

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 for several reasons. One reason is that the short-term linear prediction is segment-dependent and speaker-dependent [9]. A second reason is that by cascading short-term and long-term predictive models, only the residue contributes in the estimation the dysperiodicity via the analysis performed by the second stage with the weighted sum being discarded resulting in an estimate of the signal dysperiodicity that is different from the actual one present in the overall signal. Moreover, by inspecting the long-term prediction modeling, one can see that the estimation may lead to an inconsistency with the initial assumption of the periodicity since there is no constraint imposed to the prediction coefficients. This article proposes a new strategy for dysperiodicity estimation based on a generalized form of the variogram frequently used within the geostatistical community. Due to its ability to estimate the dysperiodicities, the generalized variogram allows to overcome the drawbacks of both long-term and short-term linear prediction formulations. The paper is organized as follows. In section 2, the long-term linear predictive model is introduced. In section 3, the generalized variogram-based approach is presented. Results based on both synthetic and real speech signals are presented in section 4. Finally, to conclude, remarks are given in section 5.

2. LONG-TERM LINEAR PREDICTION ANALYSIS

Long-term linear prediction was originally introduced in the framework of speech coding. To remove the far-sample redundancies, a long-term analysis was applied to the prediction error resulted from the conventional short-term linear prediction analysis of the speech signal [10]. Due to cycle-to-cycle prediction, the long-term predictive model allows to isolate and quantify dissimilarities between neighboring cycles. Let $x(n)$ be a stationary discrete-time zero-mean signal. The long term predictor of $x(n)$ may be expressed as

$$\hat{x}(n) = \sum_{i=0}^{m-1} a_i x(n - T - i) \quad (1)$$

where $m \geq 1$ is the order of the model; a_i, $i = 0, \ldots m-1$, are the parameters to be computed; and T is the prediction distance that is related to the vocal cycle length in the case of voiced speech sound. The order m 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. The coefficients of the model are calculated by minimizing the mean square error. The choice of the optimum value of the lag T was addressed in [10] where an exhaustive search for the optimal lag that minimizes the mean square prediction error as well as a less computationally expensive method were proposed.

Even the long-term linear predictive analysis seems to be attractive for quantifying speech signal dysperiodicity, it may result in an inconsistency with the assumption of the pseudo-periodicity of the signal. Indeed, without loss of generality, if we consider a first-order linear predictive model, solving (1) can lead to $\hat{x}(n) = -x(n-T)$ because the weighting parameter is not guaranteed to be positive.

3. VARIOGRAM-BASED SIGNAL DYSPERIODICITY ESTIMATION

3.1. Variogram

The variogram is extensively used in geostatistical data processing. In the stationary case, it is closely related to the autocovariance function. In [11], the variogram has been extended to nonstationary time series analysis. The variogram is defined as

$$\gamma(h) = \frac{1}{2} var\left(x(n+h) - x(n)\right). \quad (2)$$

For an accurate estimation of the variogram, a large number of realizations of the process is required. However, in practical situations, N samples from only one realization are available. Given a mean-stationary process, the following biased variogram estimator can be used

$$\hat{\gamma}(h) = \frac{1}{2} \frac{1}{(N-h)} \sum_{i=0}^{N-1-h} \left(x(i+h) - x(i)\right)^2. \quad (3)$$

It is worth noting that the variogram is very similar to the average magnitude difference function (AMDF) used in speech processing community, but the latter uses an absolute value.

Periodic signals are characterized as having a variogram which is zero at lags located at the period T_0 or at multiple of it and, therefore, the period may be calculated as the first lag for which the variogram is zero. This is due to the fact that if a signal is periodic, the present cycle can be perfectly estimated by means of the previous cycles. A slight perturbation of the periodic signal is interpreted as a dysperidicity which causes a small increase in the value of the variogram at lag $h = T_0$. Conversely, if the periodic signal is noise-corrupted, the minimal value of the variogram provides an estimate of the perturbation power whereas the lag associated to that minimum value may be considered as an estimate of the period of the periodic part. Thus, the variogram may be considered as a cue of the amount of the dysperiodicity (noise) present in the signal and then a SNR can be defined. However, voiced speech signals are known to be pseudo-periodic. To account for this property, some modification must be introduced in the original definition of the variogram.

3.2. Dysperiodicity Estimation by Means of the Generalized Variogram

In a typical situation of a periodic signal $x(n)$ of period T_0, we ca write $x(n) = x(n+T_0)$. Altering the signal by some random perturbation (noise) results in some dysperiodicity. The measure of the "distance from the periodicity" provides a good indication on the amount of the perturbation. Thus, a reasonable estimate of the dysperiodicity power may be

$$\varepsilon^2 = var\left(x(n+T_0) - x(n)\right). \quad (4)$$

However, voiced speech signals are pseudo-periodic in nature and characterized by smooth changes in amplitude. As an extension of the definition of the periodicity, a signal is considered pseudo-periodic if for some T_0

$$x(n+T_0) = ax(n) \quad (5)$$

where the weighting coefficient was introduced to account for amplitude changes in the speech signal. By referring to (2) and (3), and taking into account the property of pseudo-periodicity of the speech signal, the generalized variogram and its biased estimator may be defined, respectively, as

$$\gamma_g(h) = \frac{1}{2} var\left(ax(n+h) - x(n)\right) \quad (6)$$

$$\hat{\gamma}_g(h) = \frac{1}{2} \frac{1}{(N-h)} \sum_{i=0}^{N-h-1} \left(a\,x(i+h) - x(i)\right)^2. \quad (7)$$

Since the primer concern is the analysis of speech signals, the lag h is closely related to the pitch. It takes all values ranging between the minimal and the maximal fundamental periods, so that the range of the summation in the variogram estimate given by (7) may be adapted to account for this particularity. Indeed, if the range of summation is taken as is, the variogram estimate will be highly variable at higher lags due to the fewer number of lag-terms averaged. At small lags, the poorer of the variogram results from considering the fundamental frequency as unchanged on a large interval by keeping the lag constant on the overall interval which is not the case in natural speech signals. One way to avoid this problem is to take the same range in the summation for all lags. The interval must be sufficiently large to capture the dissimilarities between two neighboring cycles and sufficiently short so that the fundamental frequency can be considered constant during the analysis cycle. It has been found that an interval of 2.5 ms of length is a good choice. To take into account the nonstationarity feature of the speech signal, a forgetting factor is introduced [12]. The gain factor a is assumed to be time-dependent leading to the following expression of the generalized variogram and its estimate

$$\gamma(n,h) = \frac{1}{2} var\left(a(n)x(n+h) - x(n)\right) \quad (8)$$

$$\hat{\gamma}'_g(n,h) = \frac{1}{2L} \sum_{i=n-(L-1)}^{n} \left(a(i)\,x(i+h) - x(i)\right)^2 \quad (9)$$

where L is the summation range expressed in number of samples. For (5) to be a valid definition of the pseudo-periodicity, the

weight a must be positive. It may be interpreted as a gain factor and a reasonable choice would be

$$a(n) = \sqrt{\frac{E(n)}{E_h(n)}} \tag{10}$$

where

$$E(n) = \sum_{k=0}^{n} \lambda^{n-k} x^2(k)$$

$$E_h(n) = \sum_{k=0}^{n} \lambda^{n-k} x^2(k+h).$$

The role of the forgetting factor is to weight differently the samples. More contribution is attributed to the recent samples than the past samples. However, the forgetting factor must be less but close the unity. It fixes the effective length of the window used to calculate the gain factor. A small value of the forgetting factor may result in a short effective window whereas a value close to 1 gives rise to a long effective window. In this study, it was chosen to be $\lambda = 0.98$ which corresponds to an effective length of 2.5 *ms* at the sampling frequency f_s=20 kHz. By introducing the forgetting factor, the powers $E(n)$ and $E_h(n)$ may be expressed recursively as follows

$$E(n) = \lambda E(n-1) + x^2(n) \tag{11 - a}$$

$$E_h(n) = \lambda E_h(n-1) + x^2(n+h). \tag{11 - b}$$

The instantaneous value of the dysperiodicity for that frame is estimated as

$$e(n) = x(n) - a(n)\,x(n + h_{opt}) \qquad 0 \le n \le N-1. \tag{12}$$

The cue used as an indication on the degree of hoarseness is the SNR expressed as

$$SNR = 10 log \frac{\sum_{n=0}^{N-1} \tilde{x}^2(n)}{\sum_{n=0}^{N-1} e^2(n)} \tag{13}$$

where $\tilde{x}(n)$ is an estimate of the clean signal given by

$$\tilde{x}(n) = x(n) - e(n). \tag{14}$$

The generalized variogram must be computed in forward and backward directions. The aim of this bi-directional analysis is to remove clinically spurious errors that occur at the beginning of the record interval and at boundaries between the phonetic segments in continuous speech. It is indeed not possible to estimate 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 estimated or retro-estimated depending on which direction gives rise to the smallest variogram value. Estimation across boundaries is thus avoided and the observed error is likely to be caused by vocal perturbations rather than the evolving identity of the speech segments. Combining forward and backward variogram estimates is equivalent to vary h from $-T_{max}$ to $-T_{min}$ and from T_{min} to T_{max}, where T_{min} and T_{max} are the minimal and maximal fundamental periods in number of samples. The numerical values of T_{min} and T_{max} are 50 and 400, respectively. The algorithm is summarized as follows

1. Initialisation
   $E(0)=x^2(0)$
   $E_h(0)=x^2(h)$ for $-T_{max} \le h \le -T_{min}$ and $T_{min} \le h \le T_{max}$
2. For each time instant n=0, . . ., N-1,
   - Compute the weight $a(n)$
   - Estimate the variogram $\hat{\gamma}'_g(n,h)$ for $-T_{max} \le h \le -T_{\min}$ and $T_{\min} \le h \le T_{\max}$)
   - Find the optimal lag
   - Compute the instantaneous error $e(n)$
3. Compute the signal-to-noise ratio.

4. EXPERIMENTAL RESULTS

4.1. Data

The proposed approach has been tested on synthetic signals as well as on natural speech. Synthetic signals are very useful in judging the quality of the results and evaluating the performance of any method through some identification criteria. The artificial signals used in the test are the synthetic vowels [a] (with onset and offset) with various forms of dyperiodicities. To approximate the real analysis conditions, the sampling frequency has been chosen to be 20 kHz, the fundamental frequency has been chosen f0=120 Hz and a signal of 1 s of length has been analysed. Three sources of dysperiodicity have been considered in generating the waveforms that deviate from the perfect periodicity [2, 3]: i) additive noise, ii) dysperiodicity due to the variation in the period from cycle to cycle (jitter) and iii) dysperiodicity due to the variation in the amplitude from cycle to cycle (shimmer). For the additive noise, the SNR was varied from 8 dB to 40 dB. The amount of jitter lies between 0.1 % and 1 %, i.e., from normal to pathological case. The amount of shimmer ranges from 1 % to 10 %. For natural speech, the corpus, taken from the database in [13], is a subset of the signals corresponding to the French sentence “il est sorti avant le jour” uttered by a female speaker. 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”.

4.2. Results

Figure 1 depicts the results for the synthetic vowel [a] with the onset and offset. As can be seen, the double-prediction-based approach results in a saturation region beyond 20 dB. The actual and the variogram-based estimate SNRs are highly correlated. The superiority of the performance of the variogram-based approach appears clearly in the case of signals with jitter and shimmer. Conducted simulations have shown that the SNR values obtained by using the double prediction-based approach are noncorrelated with the amounts of jitter and shimmer. Figure 2 shows the estimate SNR versus the amount of jitter for the variogram-based approach. The variogram-based approach provides a highly correlated estimate SNR values with the amount of jitter. This is confirmed by Spearman’s rank correlation coefficient which is equal to -0.95. Figure 3 displays the estimate SNR versus the amount of shimmer. As shown, the estimate SNR is linearly decreasing as the amount of shimmer increases.

The SNR values of the different signals of continuous speech corpus are given in Table 1. The estimated SNR using the

generalized variogram 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 which corresponds to a highly dysphonic speaker is characterized by a small SNR. The double prediction-based approach provides a slightly higher SNR estimate for the signal labeled "Whisper" than for the signal labeled "rough3" resulting in an incorrect rank.

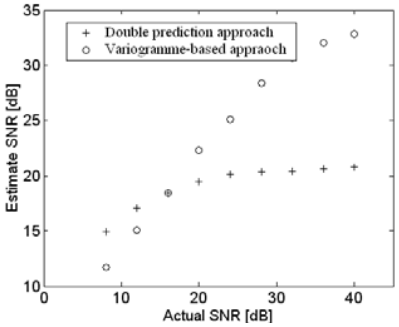

Fig. 1. Estimate SNR versus Actual SNR for the synthetic vowel [a] with onset and offset.

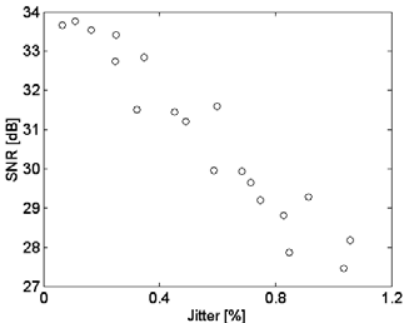

Fig. 2. Estimate SNR versus the amount of jitter in the synthetic wowel [a] with onset and offset.

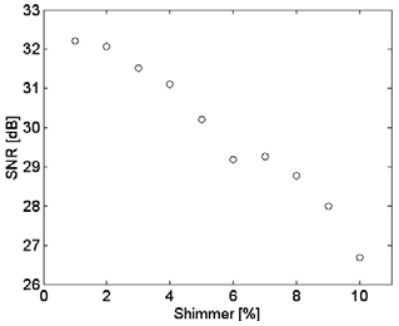

Fig. 3. Estimate SNR versus the amount of shimmer in the synthetic wowel [a] with onset and offset.

Signal	Generalized variogram	Double prediction
Modal	17.5 dB	23.4 dB
Rough1	10.5 dB	15.6 dB
Rough2	6.9 dB	12.9 dB
Rough3	3.9 dB	9.1 dB
Whisper	3.5 dB	9.2 dB

Table 1. SNR estimate of continuous speech signals using double prediction-based approach and generalized variogram.

5. CONCLUSION

This paper presents a new strategy for estimating the dyperiodicities in disordered speech. By using synthetic and natural speech signals, it has been shown that the proposed approach outperforms the double prediction-based technique. What makes this approach attractive is its ability to estimate the amount of dysperiodicity in the signal independently of its temporal or spectral structure as the conventional techniques do. The variogram-based approach requires an exhaustive search in the interval of the permissible values to get an estimate of the optimal lag. However, it is possible to reduce the computational cost either by calculating an approximate value of the optimal lag or by performing a coarse search to get an approximate value of the optimal lag followed by a fine search in the neighbouring of the approximate value to find a more accurate optimal lag.

6. REFERENCES

[1] E. Yumoto and W. J. Gould, "The estimation of signal-to-noise ratio in continuous speech of disordered voices", *J. Acoust. Soc. Am.,* vol. 71, no. 6, pp. 1544-1549, 1982.

[2] J. Schoentgen, "Spectral models of additive and modulation noise in speech and phonatory excitation signals", *J. Acoust. Soc. Am.*, vol. 113, no. 1, pp. 553-562, 2003.

[3] P. Murphy, "Spectral characterization of jitter, shimmer and additive noise in synthetically generated voice signals ", *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 978-988, 2000.

[4] H. Muta, T. Baer, K. Wagatsuma, T. Muraoka and H. Fukuda, "A pitch-synchronous analysis of hoarseness in continuous speech", *J. Acoust. Soc. Am.*, vol. 84, no. 4, pp. 1292-1301, 1988.

[5] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of continuous speech *versus* sustained vowels", *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2218-2224, 1990.

[6] J. Hillenbrand and R.A. Houde, "Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech", *J. Speech Hear. Res*., vol. 39, pp. 311-321, 1996.

[7] Y. Qi, R.E. Hillman and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech of disordered voices", *J. Acoust. Soc. Am.*, vol. 105, no. 4, pp. 2532-2535, 1999.

[8] J. Parsa, D. G- Jamieson, "Acoustic discrimination of pathological voice : Sustained vowels Versus Continuous Speech", *J. Speech Hear. Res*, vol. 44, no. 4, pp. 327-339, 2001.

[9] J. Schoentgen, "Quantitative evaluation of the discrimination performance of acoustic features in detecting laryngeal pathology", *Speech Commun.*, vol. 1, pp. 269-282, 1982.

[10] R. Ramachandran and P. Kabal, "Pitch Prediction Filters in Speech Coding", *IEEE Trans. Acoust., Speech, Sig. Proc*., vol. 37, pp. 273-331, 1989.

[11] J. Haslett, "On the sample variogram and sample autocovariance for non-stationary time series", The Statistician, vol. 46, no. 4, pp. 475-485, 1997.

[12] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall International, 2nd ed., 1991.

[13] http://www.limsi.fr/WkG/VOQUAL/