Issues in Clinical Applications of Bilateral Multi-step Predictive Analysis of Speech

J. Schoentgen, E. Dessalle, A. Kacha, and F. Grenez

National Fund for Scientific Research, Belgium
Department "Signals and Waves", Faculty of Applied Sciences,
Université Libre de Bruxelles, Brussel, Belgium

Abstract. The article concerns methodological problems posed by multi-step predictive analysis of speech, carried out with a view to estimating vocal dysperiodicities. Problems that are discussed are the following. First, the stability of the multi-step predictive synthesis filter; second, the decrease of quantization noise by means of multiple prediction coefficients; third, the implementation of multi-step predictive analyses via lattice filters; fourth, the adequacy per se of the multi-step predictive analysis paradigm for estimating vocal dysperiodicities. Results suggest that implementations of multi-step predictive analyses that are considered to be optimal for speech coding are sub-optimal for clinical applications and vice versa. Also, multi-step predictive analyses as such do not appear to be under all circumstances a paradigm adequate for analysing vocal dysperiodicities unambiguously. An alternative is discussed, which is based on a generalized variogram of the speech signal.

1 Introduction

The presentation concerns issues in clinical applications of bilateral multi-step predictive analysis of speech. Multi-step analysis designates the linear prediction of the present speech sample by means of samples that are distant. Because the purpose is the estimation of dysperiodicities in speech, the prediction distance is assigned to the lag for which the correlation between the present and a distant speech frame is maximal. This lag is indeed expected to agree with an integer multiple of the vocal cycle lengths of voiced speech sounds. In the case of unvoiced sounds or highly irregularly voiced sounds, this lag remains mathematically meaningful but is not interpreted in terms of the glottal cycle length. Bilateral means that predictive analyses are performed to the right and left of the current speech frame and that the minimal prediction error is kept and assigned to the vocal dysperiodicity trace.

Voice disorders, or dysphonias, are common consequences of disease, injury or faulty use of the larynx. A frequent symptom of dysphonia is increased noise in the speech signal or lack of regularity of the vocal cycles. Speech analyses are therefore carried out routinely in the context of the functional assessment of voice disorders.

At present, these analyses are most often carried out on steady fragments of sustained vowels. The reason is that the signal processing is often based on the

M. Faundez-Zanuy et al. (Eds.): NOLISP 2005, LNAI 3817, pp. 196 – 205, 2005.

assumption that the speech cycles are locally quasi-identical in length and amplitude. Therefore, such analyses may fail on sustained vowels or connected speech produced by severely hoarse speakers. Studies devoted to vocal dysperiodicities in connected speech or vowels including onsets and offsets are therefore comparatively rare. An overview of published research is given in [2].

Clinicians have, however, expressed the wish to be able to analyze any speech fragment produced by any speaker, including vowel onsets and offsets as well as connected speech. Arguments in favour are that, compared to stationary speech fragments, connected speech is more difficult to produce because of more frequent voice onsets and offsets, the voicing of obstruents, the maintaining of voicing while the larynx continually ascends and descends in the neck, as well as because of intonation and accentuation.

Qi et al. [1] and Bettens et al. [2] have presented methods that enable estimating vocal dysperiodicities without making any strong assumptions with regard to the regularity of the vibrations of the vocal folds or recorded speech sounds. These methods have been inspired by speech coding based on multi-step linear predictive analysis. The method presented by Qi et al. [1] involves a conventional single-step predictive analysis followed by a multi-step analysis of the residual error of the single-step prediction. In a clinical context, the multi-step prediction error is construed as the vocal dysperiodicity trace.

The method presented by Bettens et al. [2] involves a bilateral multi-step predictive analysis. It may be carried out on the speech signal directly or on any other signal considered to be clinically apposite, because the method omits the single-step analysis and avoids predicting across phonetic boundaries.

The topic of this article is an examination of methodological problems posed by bilateral multi-step predictive analyses when applied clinically.

2 Models

Formally, bilateral multi-step prediction is based on models (1). In [2], bilateral prediction is called bidirectional. In the present text, the term bilateral is preferred because it stresses the distinction between multi-step predictive analyses that are carried out to the left and right of the current speech frame, on the one hand, and the forward and backward errors involved in the lattice filter implementation of unilateral multi-step analyses, on the other.

$$e(n)=\begin{cases} e_{right}(n)=s(n)+\sum_{i=0}^{M}a_{rigth,i}s(n+T_{right}-i), \\ e_{left}(n)=s(n)+\sum_{i=0}^{M}a_{left,i}s(n-T_{left}+i). \end{cases} \tag{1}$$

Symbol *s(n)* is the current speech sample; *e(n)* is the bilateral multi-step prediction error; weights *a* are the prediction coefficients. For each analysis frame, the multi-step prediction error, the energy of which is smallest, is assigned to the dysperiodicity trace. The comparison of the present speech frame to frames to the left and right guarantees that it is compared at least once to a frame that belongs to the same phonetic segment, provided that the segment is at least two vocal cycles long. The selection of the minimum prediction error to the left or right so removes predictions that are performed across phonetic boundaries. Cross-boundary prediction errors must be discarded, because cycle-to-cycle differences owing to the evolving phonetic identity dwarf cycle discrepancies that are due to vocal noise.

Order M is typically equal to 1 or 2. The purpose of including more than one prediction coefficient is the expected reduction of quantization noise. Indeed, lag T is an integer, whereas the vocal cycle lengths are likely to be equal to a non-integer number of sampling steps. Lags T in relations (1) are determined for each analysis frame either by an exhaustive search for the minimum error or by means of the empirical inter-correlation between present and lagged frames. In the case of the latter, lag T is assigned to the position, within the open lag interval, for which the inter-correlation function is a maximum.

3 Problems and Solutions

Results show that methods proposed in [1] as well as [2] enable computing markers of vocal noise that are plausible and that co-vary with the degree of perceived hoarseness of sustained vowels or connected speech. This article is devoted to methodological issues that are raised by these proposals, as well as to their solutions.

3.1 Burg's Rule

Multi-step predictive analyses have been implemented by means of lattice filters, the coefficients of which obey Burg's rule [3]. That is, the filter coefficients are determined by means of the harmonic mean of unilateral forward and backward prediction errors, a choice that guarantees filter stability. A consequence is that the filter may be unable to track rapid signal onsets faithfully. Transients may therefore give rise to prediction errors that are higher than the prediction errors that one would obtain by means of unstable filters.

Owing to the bilateral analysis, however, this is likely to be a problem only when a rapid signal boost ends or a rapid signal drop starts at a phonetic boundary. When no risk of cross-boundary prediction is involved, the bilateral analysis turns the prediction of onsets into the retro-diction of offsets and vice versa.

Be that as it may, in the framework of clinical applications linear multi-step prediction is carried out for analysis purposes only. Filter stability is therefore not an issue and can be omitted in favour of a direct form implementation the coefficients of which are determined by means of the conventional covariance method, for instance.

3.2 Lattice Filter Implementation

When more than one multi-step prediction coefficient is involved, the prediction error obtained by a lattice filter comprises several recent as well as several distant speech samples. For instance, when order M is equal to 1, the lattice filter output is the following [3].

$$e_{left}(n) = s(n) + c_T c_{T-1} s(n-1) + c_{T-1} s(n-T+1) + c_T s(n-T). \tag{2}$$

Symbol *T* is the prediction distance in number of samples, *s(n)* is the *nth* speech sample and c_j are the lattice-filter coefficients. Sample *s(n-1)* in error (2) obscures the conceptual simplicity of relations (1) and upsets the straightforward interpretation of the multi-step prediction error as a measure of vocal dysperiodicity. The intercalation of additional recent samples is typical of the lattice filter implementation and can be avoided in the framework of implementations that are direct or involve single coefficients only.

3.3 Multiple Prediction Coefficients

Relations (1) may involve multiple prediction coefficients. A consequence is that the present speech sample is compared to a weighted sum of distant speech samples. The goal is to decrease quantization noise. A sample-by-sample comparison by means of a single-coefficient multi-step prediction would be easier to interpret, however, given the overall objective, which is to estimate vocal dysperiodicities.

A solution consists in decreasing quantization noise by over-sampling first and replacing the multiple coefficients by a single one. This removes the risk of decreasing genuine vocal noise via the weighted sum that is involved in the distant prediction.

3.4 Multi-step Linear Predictive Analysis as a Paradigm for the Analysis of Vocal Dysperiodicities

This section addresses a basic issue, which is the adequacy per se of the multi-step prediction paradigm as a framework for analyzing vocal dysperiodicities. Hereafter, one assumes that the multi-step prediction involves a single coefficient the value of which is determined by means of the conventional covariance method. The conclusions are valid, however, for any implementation of the multi-step predictive analysis filter.

The covariance method consists in minimizing the energy of the prediction error cumulated over a rectangular frame of length *N*. When a single coefficient is involved, one easily shows that the (unilateral) multi-step prediction error is equal to the following.

$$E=\sum_{n=1}^{N}\left[s(n)-s(n-T)\frac{\sum_{i=1}^{N}s(i)s(i-T)}{\sum_{i=1}^{N}s^2(i-T)}\right]^2 . \tag{3}$$

From (3) follow solutions (4). Parameter *b* is a positive gain that is constant over the analysis frame. It demonstrates that the prediction coefficient in (3) automatically compensates for slow variations of the vocal amplitude.

$$E=0\ \begin{cases} s(n)=+bs(n-T), \\ s(n)=-bs(n-T). \end{cases} \tag{4}$$

Solutions (4) show that, formally, the multi-step prediction error is not a measure of vocal dysperiodicity. The reason is parasitic solution *s(n) =-bs(n-T)*. For a sinusoid of period *T*, for instance, solutions (4) suggest that the multi-step prediction error is a minimum for shifts *T/2* and *T*, of which only the latter has an interpretation in terms of the period of the sinusoid. In practice, this means that an exhaustive search for optimal shift *T* is likely to produce erroneous measures of dysperiodicity for phonetic segments that are quasi-sinusoidal, i.e. voiced plosives, for example.

Determining optimal shift *T* by means of the empirical inter-correlation between present and lagged frames is less likely to give rise to parasitic solutions. The reason is that the optimal shift is assigned to the lag for which the inter-correlation is a maximum. Formally, the removal of parasitic solutions is not guaranteed, however.

Moreover, the interpretation of error *E* remains ambiguous even when parasitic solutions are discarded. Because of the inter-correlation that is involved in (3), error *E* is a measure of signal dysperiodicity only when the vocal noise is feeble. The prediction error turns into a measure of signal energy when the vocal noise is strong (Table 1).

3.5 Generalized Variogram

A possible alternative is based on the observation that for a periodic signal *s(n)*, the following expression is expected to be true for any shift *T* that is an integer multiple of the signal period, assuming that the quantization noise can be neglected.

$$\sum_{n=-\infty}^{n=+\infty}\left[s(n)-s(n-T)\right]=0. \tag{5}$$

In practice, voiced speech segments are locally-periodic at best, speech cycle amplitudes are expected to evolve slowly and the glottal cycle length is not known a priori. This suggests analyzing the signal frame by frame, squaring expression (5), and inserting a positive gain *g* that is constant over the analysis frame.

$$V(T)=\sum_{n=0}^{N}\left[s(n)-gs(n-T)\right]^2. \tag{6}$$

When gain *g* = 1, cumulated difference (6) is known as the empirical variogram of signal *s(n).* Length *N* fixes the frame length. The squaring guarantees that difference (6) is a minimum for lags that are integer multiples of the period of the signal.

Gain *g* enables neutralizing drifts of the signal amplitude that are due to onsets, offsets or prosody. Gain *g* is chosen so that it is always positive and the interpretation of generalized variogram *V(T)* is the same whatever the strength of the vocal noise. A definition of *g* that satisfies these criteria equalizes the signal energies in the present and lagged analysis frames.

$$V(T)=\sum_{n=0}^{N}\left[s(n)-s(n-T)\sqrt{\frac{\sum_{i=0}^{N}s^2(i)}{\sum_{i=0}^{N}s^2(i-T)}}\right]^2. \tag{7}$$

Table 1. Variogram (7) and multi-step prediction error (3) for periodic, odd-periodic and white noise signals

	white noise	$s(n) = -bs(n-T), b > 0$	$s(n) = bs(n-T), b > 0$
V	$\Sigma[s(n)-s(n-T)]^2$	$4\Sigma s^2(n)$	0
E	$\Sigma s^2(n)$	0	0

Inspecting multi-step prediction error (3) and generalized variogram (7) suggests that they are proportional when $s(n)$ is approximately equal to $s(n-T)$. Otherwise, they are different. Table 1 summarizes the values of expressions (3) and (7) when, for example, $s(n) = bs(n-T)$, $s(n) = -bs(n-T)$, $b > 0$, as well as when $s(n)$ is white noise.

One sees that generalized variogram V is different from zero when the signal is odd-periodic and lag T equal to the odd-period. Also, expression V is the cumulated squared difference between the present and lagged signal samples, whether the signal is deterministic or stochastic. The minimum of V is therefore a measure of signal dysperiodicity in the analysis frame.

On the contrary, the multi-step prediction error E is zero when the signal is periodic or odd-periodic and lag T equal to the period or odd-period. Also, error E is the cumulated squared difference between the present and lagged signal samples only when they are (strongly) correlated. When they are uncorrelated, error E is the signal energy. Error E is therefore a measure of signal (un)-predictability. Because predictability is a more general property than periodicity, variogram V and error E only agree for special instances of signals and lags.

4 Methods

The experimental part of the study involves seven analysis methods, which are listed in Table 2. The objective is to investigate whether issues that are discussed above give rise to statistically significant differences in the vocal dysperiodicity traces. For each method, the length of the rectangular analysis frame was equal to 2.5 milliseconds [2]. The analysis frames were non-overlapping, but contiguous. Prediction lag T was assigned to the position of the maximum of the inter-correlation between present and lagged frames or, when appropriate, to the position of the minimum of the variogram. The prediction lag was requested to be within an interval between 2.5 and 20 milliseconds. This interval includes the phonatory cycle lengths that are typical of male and female speakers. Per frame, each analysis method was applied twice, once for positive and once for negative lag values, and the minimum prediction error or variogram-determined signal difference was kept and assigned to the vocal dysperiodicity trace.

For several analyses, the speech signals, inter-correlation function or variogram were interpolated linearly or parabolically. The purpose was to test the use of non-integer prediction lags.

4.1 Analysis Methods

Table 2. Characteristics of analysis methods

Label	Analysis method	Nber of coefficients	Interpolation
1	Burg, covariance-lattice	3	no
2	covariance	1	no
3	covariance	3	no
4	covariance	1	linear
5	covariance	1	parabolic
6	variogram	n.a.	no
7	variogram	n.a.	linear

4.2 Corpora

The corpora have been sinusoids; as well as vowels and short sentences produced by normophonic or dysphonic speakers. Sinusoids as well as speech signals have been sampled at 20 kHz. The sinusoids have been contaminated by additive or frequency modulation noise. The purpose was to test interpolation with a view to reducing quantization noise.

The speech corpus comprised sustained vowels [a] and two French sentences spoken affirmatively by 22 normophonic or dysphonic, male or female speakers. The sentences were “le garde a endigué l’abbé” (S1) and “une poule a picoré ton cake” (S2). All phonetic segments in sentence S1 are voiced by default, whereas sentence S2 comprises voiced as well as unvoiced phonetic segments. The sentences are matched grammatically and comprise the same number of syllables. Strident fricatives were omitted on purpose.

4.3 Noise Marker

The vocal dysperiodicity trace *e(n)* is summarized by means of a signal-to-dysperiodicity ratio (*SDR*) that is defined as follows [1]. Symbol *I* is the number of samples in the total analysis interval.

$$SDR = 10\log\frac{\sum_{i=1}^{I} s^2(i)}{\sum_{i=1}^{I} e^2(i)}. \tag{7}$$

Table 1 shows that *SDR* → 0 when the signal is white noise and analyzed by means of multi-step prediction. On the contrary, *SDR* → -3 dB when the signal is white noise and analyzed by means of the generalized variogram. The reason is that variogram (6) is the cumulated squared difference between present and lagged samples. Prediction error (3) is, on the contrary, equal to a cumulated squared difference between present and lagged samples only when the signal is periodic or pseudo-periodic.

5 Results and Discussion

5.1 Sinusoidal Signals

Analyses of sinusoids confirm that dysperiodicity traces obtained by single-coefficient multi-step predictive or variogram analyses may be altered by quantization noise. *SDR* values of clean sinusoids sampled at 20 kHz were typically comprised in the interval 30 – 40 dB when the sampling frequency was not an integer multiple of the frequency of the sinusoid.

Non-integer lags, determined via interpolation, have been shown to increase the distance between vocal and quantization noise. Simulations suggest that interpolation moves the *SDR* values of sampled clean sinusoids to values greater than 65 dB.

5.2 Sustained Vowels and Running Speech

Table 3 summarizes the quartiles of the *SDR* values (in dB) obtained for a corpus of sustained vowels [a], including onsets and offsets, and sentences *S1* and *S2* spoken by 22 speakers. The labels of the analysis methods agree with the labels given in Table 2. The *SDR* values have been rounded to the nearest decimal after the comma.

For each speech corpus, a single-factor repeated measures analysis of variance of the *SDR* values has been carried out to check whether differences between methods 1 to 7 are statistically significant. Subsequently, methods have been compared pair-wise by means of paired *t*-tests. The levels of significance of the individual tests have been adjusted by means of Bonferroni's correction to fix to *0.05* the overall level of significance of a total of *21* pair-wise comparisons [5]. Statistical analyses of the data show the following.

a) For vowel [a], the analysis of variance shows that the inter-method differences are statistically significant ($F = 249$, $p < 0.001$). Out of the *21* pair- wise comparisons, *17* are statistically significant. Of these, all involve differences between analysis methods (covariance-lattice, covariance of order 0 or 2, variogram).

Table 3. Quartiles of the *SDR* values (in dB) obtained for a corpus of sustained vowels [a], including onsets and offsets, and sentences *S1* and *S2* spoken by 22 speakers

	Method label	1	2	3	4	5	6	7
[a]	First quartile	23.5	16.8	17.0	16.7	16.7	16.7	16.7
	Median	26.7	20.2	20.6	20.3	20.4	20.1	20.2
	Third quartile	28.7	22.4	22.9	22.6	22.8	22.4	22.6
S1	First quartile	19.5	14.4	14.6	14.4	14.4	14.2	14.2
	Median	22.3	17.2	17.4	17.2	17.1	17.2	17.1
	Third quartile	24.6	18.1	18.4	18.1	17.9	18.0	18.0
S2	First quartile	19.0	16.7	17.2	16.6	15.7	16.5	16.8
	Median	22.6	18.3	18.5	18.2	17.7	18.1	18.1
	Third quartile	24.4	19.7	20.0	19.4	19.2	19.6	19.4

b) For sentence *S1*, the analysis of variance shows that the inter-method differences are statistically significant (*F = 129*, *p < 0.001*). Of the *21* pair-wise comparisons, *16* are statistically significant. Of these, *15* pairs involve differences between analysis methods (covariance-lattice, covariance of order 0 or 2, variogram). One pair differs by the interpolation method (linear versus parabolic).
c) For sentence *S2*, the analysis of variance shows that the inter-method differences are statistically significant (*F = 67*, *p < 0.001*). Of the *21* pair-wise comparisons, *15* are statistically significant. Of these, all involve differences between analysis methods (covariance-lattice, covariance of order 0 or 2, variogram).

Results therefore suggest that different analysis methods cause *SDR* values to differ statistically significantly. Possible explanations are the following.

a) The covariance-lattice implementation (Table 3, column 1) implicates running averages of the recent as well as distant samples. The original purpose of involving several prediction coefficients has been the decrease of quantization noise. Results suggest that multiple prediction coefficients decrease genuine vocal noise as well as quantization noise.

   Also, the lattice filter is stable. Stability would let one expect a boost of the prediction error because of an increased difficulty in tracking rapid transients. This is not observed. This would suggest that either the corresponding error increase is masked by the decrease of genuine vocal noise owing to local averaging (2), or by the bilateral analysis (1) that turns onsets into offsets.
b) The 3-coefficient covariance method (Table 3, column 3) involves a running average of the distant samples only. The original purpose has been the decrease of quantization noise. Single-coefficient covariance analyses omit this local smoothing. As a consequence, single-coefficient (column 2) and 3-coefficient (column 3) covariance analyses give rise to *SDR* values that differ statistically significantly. Inspecting data averages suggests that the corresponding *SDR* values typically differ by less than 1 dB. The difference is due to a decrease of the genuine vocal noise by local averaging rather than to a decrease of the quantization noise.
c) The variogram (Table 3, columns 6 and 7) involves an energy-normalisation coefficient the mathematical properties of which differ from those of the prediction coefficients implicated in methods labelled 1 to 5. Consequently, *SDR* values obtained by variogram and linear predictive analyses differ statistically significantly. Inspection of the data averages suggests, however, that *SDR* values obtained via 1-coefficient covariance and variogram analyses typically differ by less than 1 dB. Simulations indeed show that variogram and 1-coefficient linear predictive analyses give comparable *SDR* values as long as these are greater than roughly 10 dB [4].

Statistical analyses show that interpolation does not cause the *SDR* values to increase statistically significantly for a same analysis method. The purpose of interpolation is to decrease quantization noise. Inspecting data averages suggests that *SDR* differences owing to interpolation are typically less than 0.1 dB. A possible explanation is that, in the absence of interpolation, the *SDR* ceiling owing to quantization noise is in the vicinity of 30 dB. Therefore, quantization noise is negligible compared to vocal noise in signals the *SDR* value of which is typically 17 dB.

6 Conclusion

Implementations of linear predictive analyses that are considered to be optimal for speech coding are sub-optimal for clinical applications and vice versa. For clinical applications, the recommended implementation would involve a single prediction coefficient the value of which is fixed by means of a conventional covariance method. Interpolation or over-sampling would be the preferred method for decreasing quantization noise. Moreover, the presentation shows that multi-step prediction is not a paradigm that would enable interpreting under all circumstances the prediction error as a trace of the vocal dysperiodicity. The generalized variogram of the speech signal is an alternative that does not admit any ambiguity in interpretation.

References

[1] Qi Y., Hillman R. E., and Milstein C. (1999) ‘‘The estimation of signal to-noise ratio in continuous speech for disordered voices,’’J.Acoust. Soc. Am. 105, 4, 2532–2535.

[2] Bettens F., Grenez F. and Schoentgen J. (2005) “Estimation of vocal dysperiodicities in disordered connected speech by means of distant-sample bidirectional linear predictive analysis”, J. Acoust. Soc Am., 117, 1, 10 pp.

[3] Ramachandran R., and Kabal P. (1989) ‘‘Pitch prediction filters in speech coding,’’IEEETrans. Acoust., Speech, Signal Process. 37, 4, 467–478.

[4] Dessalle, E. (2004) “Estimation en ligne des dispériodicités vocals dans la parole connectée”, unpublished Master Thesis, Université Libre de Bruxelles, Bruxelles.

[5] Moore D., McCabe G. (1999) “Introduction to the practice of statistics”, Freeman, New York.