Assessment of vocal dysperiodicities in connected disordered speech

A. Alpan [1], *A. Kacha* [1], *F. Grenez* [1], *J. Schoentgen* [1,2]

[1] Laboratoire d'Images, Signaux et Dispositifs de Télécommunications,
Université Libre de Bruxelles, Brussels, Belgium
[2] National Fund for Scientific Research, Belgium

aalpan@ulb.ac.be, akacha@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be

10.21437/Interspeech.2007-381

Abstract

The aim of the presentation is to investigate acoustic analysis of connected speech by means of an average-equalized and energy-equalized variogram to extract vocal dysperiodicities. The variogram enables positioning a current and a lagged analysis frame in adjacent speech cycles to track inter-cycle dysperiodicities. Average and energy equalization of the analysis frames are options that make it possible to compensate for slow deterministic changes of the speech signal amplitude in connected speech. The instantaneous dysperiodicity trace has been summarized by means of segmental and global signal-to-dysperiodicity ratios. Results show that signal-to-dysperiodicity ratios obtained by variogram analysis correlate strongly with the perceived degree of hoarseness when the analysis frames are energy-equalized. Equalizing the frame averages removes small artifacts in the instantaneous dysperiodicity trace that are caused by sound-to-sound transients or intrusive low-frequency noise.

Index Terms: connected disordered speech analysis, variogram, signal-to-dysperiodicity ratio.

1. Introduction

The presentation concerns the problem of tracking vocal dysperiodicities in connected speech, with a view to characterizing voice disorders. Acoustic analysis of speech is non-invasive and enables clinicians to monitor and express numerically the degree of hoarseness of a speaker's voice.

Most acoustic clinically relevant cues are obtained from steady fragments of sustained vowels. The use of sustained vowels is the consequence of technical feasibility rather than clinical relevance. For sustained sounds, the hypotheses of stationarity and pseudo-periodicity are valid for many speakers, if onsets and offsets are omitted [1].

However, clinicians consider connected speech to be more informative than sustained vowels. This is a matter of clinical experience. Moreover, the perceptual evaluation of voice is mainly based on connected speech or sustained vowels including onsets and offsets, rather that on steady fragments of vowels. Connected speech or onsets and offsets indeed inform about the transient characteristics of vocal fold vibration.

Several acoustic cues have been used to characterize speech of dysphonic patients. The most popular describe the deviation of the speech signal from strict periodicity [5,6]. The causes of observed disturbances are multiple. They include speech cycle jitter, signal amplitude shimmer, and additive noise owing to turbulence, diplophonia, random vocal fold vibration, and parasitic vibrations of the false vocal folds, as well as uncontrolled transients between different vocal regimes. Existing acoustic signal analysis techniques therefore aim, more often than not, at the isolation of individual cycles in the speech signal or harmonics in the speech spectrum [1]. This task may be error-prone or impossible for speech sounds produced by severely hoarse speakers, and a consequence is that these techniques may be not reliable [7].

In [2], Kacha et al. have proposed a generalized variogram to estimate vocal dysperiodicities. The generalized variogram enables tracking vocal dysperiodicities in running speech without an a priori knowledge of the average cycle length. It is based on the comparison of two analysis frames that are positioned in neighbouring speech cycles. The method takes into account the deterministic evolution of the signal amplitude owing to sound onsets and offsets, or sound-specific intensities by equalizing the signal energy across analysis frames. The signal average has, however, been assumed to be constant.

But, it appears that this assumption is not valid under all circumstances. Violations of that hypothesis are observed in transients when the signal shape evolves along with the signal amplitude. Violations also occur because of low-frequency noise, examples of which are pop noise due to the speaker's breath hitting the recording microphone that has not been positioned carefully. Hereafter, it is shown that equalizing averages across analysis frames enables improving the correlation between computed signal-to-dysperiodicity ratios and perceptual scores of hoarseness, as well as discarding parasitic low-frequency transients.

2. Methods

2.1. Generalized variogram analysis [2]

For a periodic signal $x(n)$ of period T_0, one may write $x(n) = x(n-T_0)$. For stationary signals, the deviation from strict periodicity over an analysis frame of length N can be estimated by the following expression.

$$\delta = \min_{T} \left\{ \sum_{n=0}^{N-1} \left[x(n) - x(n+T) \right]^2 \right\}, -T_{\max} < T < -T_{\min},\ T_{\min} < T < T_{\max} \qquad (1)$$

The expression between accolades in (1) is known as the variogram of the speech signal. The expression between brackets equals the difference between a current and a lagged analysis frame of length N. Index n positions the samples within the frame. In practice, lag T is allowed to vary between ± 2.5 and ± 20 *ms*. Lag T may be positive as well as negative. This is an option that guarantees the comparison of intra-sound cycles exclusively and avoids comparing cycles across phonetic boundaries. Indeed, when a speech cycle is near the right-hand boundary of a phonetic segment the segment-internal cycles are expected to be to its left, that is, lag T is expected to be negative. In expression (1), delay T is fixed so

as to minimize the cumulated differences between current and lagged frames. For voiced sounds, lag T is therefore an integer multiple of the glottal cycle length. In unvoiced sounds, (1) can be meaningfully computed, the interpretation of lag T in terms of glottal cycle lengths is not valid anymore, however.

In running speech, the signal amplitude evolves deterministically owing to onsets and offsets, segment-specific loudness and accentuation. To take into account these slow variations, a gain α is introduced in expression (1) to equalize the energies between the analysis frames.

$$\delta = \min_{T} \left\{ \sum_{n=0}^{N-1} \left[x(n) - \alpha x(n+T) \right]^2 \right\}, \quad \alpha = \sqrt{\frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} x^2(n+T)}} \tag{2}$$

The analysis frame length equals 2.5 *ms*. It guarantees that current and lagged frames do not overlap, because the analysis frame length is equal to the shortest permitted lag. Successive analysis frames are shifted by 2.5 *ms*, which enables computing an instantaneous frame-by-frame dysperiodicity (3).

$$e(n) = x(n) - \alpha x(n + T_{opt}) \tag{3}$$

Lag T_{opt} is the positive or negative shift that minimizes cumulated difference (2). Hereafter, energy-equalized variogram (2) is called the generalized variogram to make the distinction with the conventional variogram (1) and the average-equalized generalized variogram that is discussed hereafter.

2.2. Average-equalized generalized variogram

In the average-equalized generalized variogram, one subtracts the frame average from each speech sample before computing the energy-equalized variogram. The prior equalization of the averages has consequences for the estimation of the cumulated differences in situations where a running average over an analysis frame evolves from one cycle to the next. Frame averages are expected to evolve during segment onsets and offsets, for instance, as well as owing to clinically relevant or intrusive low-frequency noise (e.g. vocal amplitude tremor and pop noise).

2.3. Global and segmental signal-to-dysperiodicity ratios

The global signal-to-dysperiodicity ratio in dB has been computed as follows. It summarizes the relative amount of noise in a speech recording of length L.

$$SDR = 10 \log \frac{\sum_{n=0}^{L-1} x^2(n)}{\sum_{n=0}^{L-1} e^2(n)} \tag{4}$$

The segmental signal-to-dysperiodicity ratio *SDRSEG* consists in computing ratio (4) locally over intervals of 5 *ms* and then taking the average. In the framework of the evaluation of lossy speech coders, the segmental signal-to-dysperiodicity ratio is recommended over the global one, because it is known to correlate better with human-assigned scores of perceived quality. The explanation appears to be that segmental ratios boost weak noisy segments, which seem to influence perceived timbre strongly [4].

2.4. Corpus

The corpus comprises sustained vowel [a] and four French sentences produced by 22 normophonic or dysphonic speakers (10 male and 12 female speakers. The sentences have been the following: “le garde a endigué l’abbé” (S1); “une poule a picoré ton cake” (S2); “Bob m’avait guide vers les digues” (S3); “ta tante a appaté une carpe” (S4). The sampling frequency has been 48 kHz. The perceived degree of hoarseness has been determined by comparative judgments of pairs of speech tokens [3].

3. Results and discussion

Table 1 shows the Pearson correlation coefficients between perceived degree of hoarseness and global and segmental signal-to-dysperiodicity ratios. One sees that the average-equalized generalized variogram increases the correlation with the global signal-to-dysperiodicity ratio for sentences S2 to S4 by 15%, 24% and 14% respectively. For sentence S1 no increase is observed.

Table 1. Pearson’s correlation coefficients between average hoarseness scores and *SDR* and *SDRSEG* for sustained vowel [a] and sentences *S1-S4* for generalized variogram (GV) and average-equalized generalized variogram (AGV)

		[a]	S1	S2	S3	S4
GV	SDR	-0.73	-0.71	-0.68	-0.70	-0.69
	SDRSEG	-0.70	-0.85	-0,79	-0.80	-0.66
AGV	SDR	-0.70	-0.72	-0.78	-0.87	-0.79
	SDRSEG	-0.68	-0.84	-0.78	-0.82	-0.70

Replacing the generalized by the average-equalized generalized variogram analysis does not increase the correlation between the segmental signal-to-dysperiodicity ratio and perceptual scores. It is the case, however, that the correlation has been larger for the segmental *SDR*s before equalizing the averages. For sustained vowel [a] the correlation between dysperiodicities and perceptual scores is equivalent for the two ratios and the two analysis methods.

Table 1 suggests that computing the segmental signal-to-dysperiodicity ratios or equalizing the averages of the analysis frames increases the correlation between acoustic cues and perceptual scores in the case of running speech. Hereafter, we discuss possible explanations for the observed differences between analysis methods and dysperiodicity cues.

Figure 1 is a scattergram that shows for sentence S3 perceptual scores of hoarseness on the horizontal axis and the global signal-to-dysperiodicity ratio on the vertical axis, computed by means of the average-equalized and un-equalized generalized variograms. Generally speaking, the effect of equalizing frame averages in addition to frame energies has been to improve the linearity between perceptual and acoustic cues and to increase the Pearson correlation coefficient, which is a measure of linear correspondence. One sees that the difference between the two analysis methods increases with the signal-to-dysperiodicity ratio. That is, when the frame averages are not equalized, the cleaner the signal,

the better one observes small artifacts due to comparisons during transients of frames that belong to adjacent cycles.

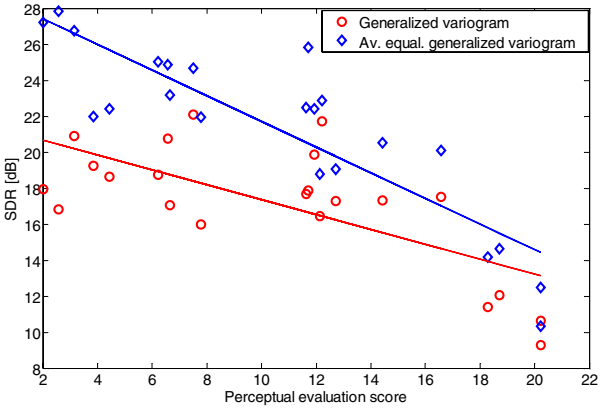

Figure 1: Global signal-to-dysperiodicity ratio (vertical axis) versus perceptual scores (horizontal axis) and linear regression lines for sentence *S3*. Increasing scores to the right on the horizontal axis correspond to increasing scores of perceived hoarseness, that is, decreasing signal-to-dysperiodicity ratios

When the segmental signal-to-dysperiodicity ratios are substituted for the global ones, the correlation increases and the difference between the average-equalized and un-equalized analyses decreases. This would suggest that isolated but high-energy transients have contributed to the measured dysperiodicity ratios of the clean signals. It is indeed the case that segmental signal-to-dysperiodicity ratios diminish the contribution of high-amplitude signal fragments and boost the contribution of low-amplitude ones.

One may indeed understand better where the differences between analysis methods in Figure 1 come from, by observing the speech waveform of sentence S3 produced by a normophonic speaker. Figure 2 represents from top to bottom the speech waveform, instantaneous dysperiodicities (3) obtained by generalized (black) and average-equalized generalized (red) variograms and the local signal-to-dysperiodicity ratios calculated over segments 5 *ms* long.

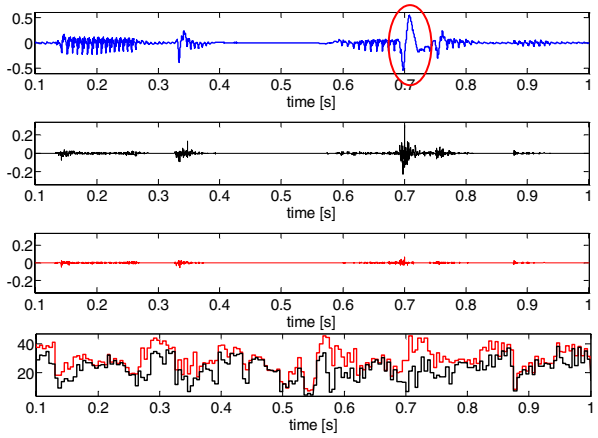

Figure 2: Speech signal, instantaneous dysperiodicity for generalized and average-equalized variogram and local SDR traces for a fragment ("Bob m'avait") of sentence S3 produced by a normophonic speaker; black and red: generalized and average-equalized generalized variograms

One notices that local signal-to-dysperiodicity ratios are larger for the average-equalized generalized variogram. This would suggest that noise is tracked better. The rest of the paragraph is devoted to the discussion of the high-amplitude transient that is highlighted in the speech waveform. This transient contains half of the signal energy. One observes high-energy low-frequency components below 50 *Hz* in its magnitude spectrum.

In the generalized variogram analysis this transient gives rise to large instantaneous dysperiodicities because it is high-amplitude and low-frequency. Even if it had repetitive features, these could not be exploited because lag T in (2) cannot exceed 20 *ms*. Because the transient only evolves slowly, the average-equalized generalized variogram decreases, however, the contribution of the transient to the dysperiodicity trace, owing to the equalization of the frame averages.

Figure 1 would suggest that the removal of low-frequency pops from the dysperiodicity trace is legitimate. This is confirmed by listening to the sentence displayed in Figure 2, which indicates that the high-amplitude popping noise is barely audible. Its timbre suggests that it is caused by breath hitting the microphone housing. Even if it were clearly audible, listeners would dismiss it because its unusual timbre and position mark it as intrusive.

Figure 3 illustrates why artifacts in the dysperiodicity trace, owing to transients, have been an issue in clean or quasi-clean signals only. Figure 3 shows the same plots as in Figure 2, but for a severely hoarse speaker. Here the intrinsic noise is so large that it masks in the dysperiodicity trace any artifacts owing to transients (whatever their cause). This is confirmed by the local signal-to-dysperiodicity ratios, which follow each other closely for the average-equalized and not-equalized analysis methods (bottom plot of Figure 3).

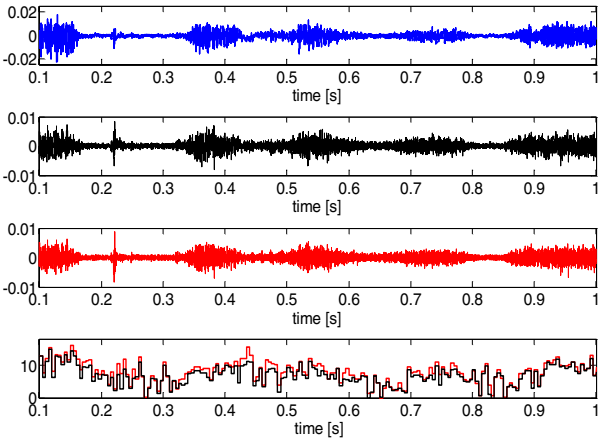

Figure 3: Signal, dysperiodicity and local SDR traces calculated for a fragment of sentence S3 produced by a severely hoarse speaker

To examine further the contribution of transient-related artifacts to the dysperiodicity trace, and their removal, simulation studies have been carried out. An artificial periodic stimulus has been formed by means of two sinusoids, with a fundamental frequency equal to 119Hz. To simulate natural low-frequency noise, the sentence displayed in Figure 2 has been low-pass filtered at a cut-off frequency of 45Hz. The low-frequency noise has then been added to the periodic signal. The energy of this low-frequency noise has been chosen to be equal to the energy of the clean periodic signal,

because it is observed in Figure 2 that in extreme cases low-frequency noise energy might be of the same order as the signal energy. Finally, white noise has been added to the sum. The signal to noise ratio between the periodic signal and the white noise has been varied from 4dB to 52dB. The global signal-to-dysperiodicity ratios obtained for both analyses methods are displayed in Figure 4.

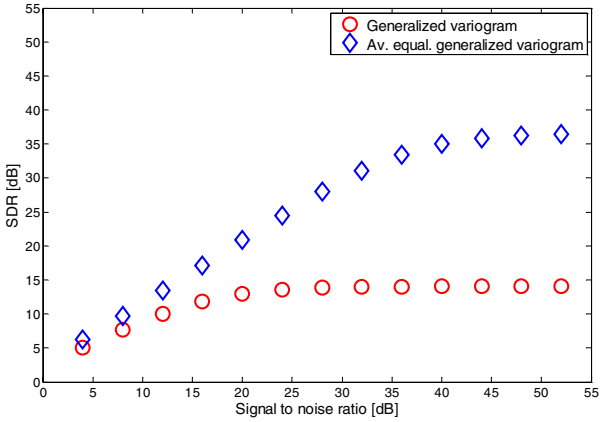

Figure 4: Global signal-to-dysperiodicity ratio as a function of the signal-to-noise ratio in the presence of large fixed low-frequency pop noise.

One sees that for small signal-to-noise ratios average-equalized and un-equalized generalized variograms give comparable global SDR values. When the signal-to-noise ratios increase, a plateau appears for the global signal-to-dysperiodicity ratio obtained via the un-equalized generalized variogram (at 15 dB). For the global signal-to-dysperiodicity ratio obtained via the average-equalized variogram a plateau is observed for higher SNR values (at 40 dB).

One should note that the low-frequency noise has been deliberately chosen to be large compared to the signal. Also, the global rather than the segmental signal-to-dysperiodicity ratios have been displayed. These choices boost the differences between the average-equalized and un-equalized generalized variograms. The choices have been made for demonstration purposes. In practice, the plateau appears for larger signal-to-noise ratios and the differences between the two analysis methods are less marked than in Figure 4. Also, the segmental signal-to-dysperiodicity ratio is less sensitive to isolated high-amplitude events than the global ratio.

A final simulation has dealt with transients that are not caused by intrusive events, but are intrinsic to the onsets and offsets of phonetic segments. The simulations comprised synthetic vowel timbres [u] including onsets and offsets. Amplitude-modulated white noise had been added to the periodic glottal source.

The signal-to-noise ratio displayed on the horizontal axis in Figure 5 is the ratio of the noise and glottal source energies. The measured global signal-to-dysperiodicity ratios are displayed on the vertical axis. One observes a divergence of the ratios obtained via the average-equalized and un-equalized generalized variograms starting at 25 dB. A plateau then follows. The explanation is that the frame average evolves in the onset and offsets. This evolution has been incompletely compensated for by energy equalization of the current and shifted analysis frames, but it appears to have been compensated by energy as well as average equalization. The plateau then is moved to 40 dB, which is way above signal-to-noise ratios expected in clean speech.

The differences between analysis methods are more easily observed in timbre [u] than in other timbres. The explanation is that vowel [u] is characterized by two low formant frequencies, which boost the differences between the frame averages in transients. Finally one should note that, onset and offset-related artifacts could not be compensated anymore by substituting segmental for global signal-to-dysperiodicity ratios. The reason is that onsets and offsets are low-amplitude compared to the vowel core. Their contribution to the dysperiodicity trace is therefore boosted in segmental signal-to-dysperiodicity ratios.

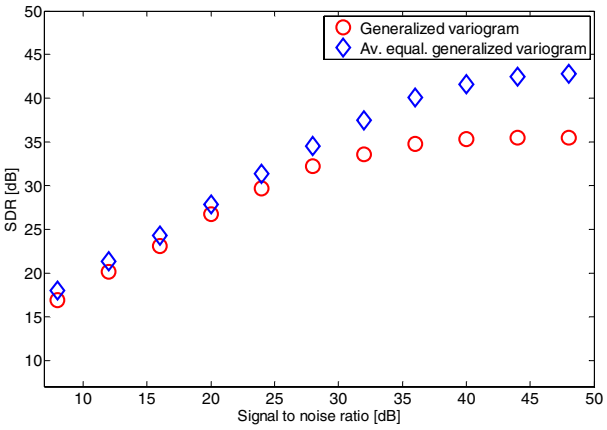

Figure 5: Global signal-to-dysperiodicity versus signal-to-noise ratio for synthetic [u] including onsets and offsets, with additive white noise at the source.

4. Acknowledgements

This research was supported by the “Région Wallonne”, Belgium, in the framework of the “WALEO II” programme.

5. References

[1] Klingholtz, F., “Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels”, J. Acoust. Soc. Amer., 87(5), pp. 2218-2224, 1990.

[2] Kacha, A., Grenez, F., and Schoentgen, J., “Estimation of dysperiodicities in disordered speech”, Speech Com., Vol. 48, pp.1365-1378, 2006.

[3] Kacha, A., Grenez, F., and Schoentgen, J., “Voice quality assessment by means of comparative judgments of speech tokens”, in: Proc. Int. Conf. Spoken Lang. Process., Lisboa, Portugal, pp.1733-1736, September 2005.

[4] N.S. Jayant and P. Noll, “Coding of waveforms: principles and applications to speech and video”, Prentice-Hall, Englewood Cliffs,1984

[5] Y. Qi, R.E. Hillman, C. Milstein, “The estimation of signal-to-noise ratio in continuous speech for disordered voices”, J. Acoust. Soc. Am., 105, 4, 2532-2535, 1999

[6] V. Parsa, D. G. Jamieson, K. Stenning, H. A. Leeper, “On the estimation of signal-to-noise ratio in continuous speech for abnormal voices”, Proc. Int. Conf. Spoken Language Processing, Interspeech 2002, Denver, 2505-2508, 2002

[7] D. L. Streiner, G. R. Norman, “Health measurement scales”, Oxford University Press, Oxford, 1995