FRAME-BASED ACOUSTIC CUES OF VOCAL DYSPERIODICITY IN CONNECTED SPEECH

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

[(1)] Department Waves and Signals, Université Libre de Bruxelles, Brussels, Belgium
[(2)] National Fund for Scientific Research, Belgium
E-mail: akacha@ulb.ac.be

ABSTRACT

Acoustic analysis of connected speech is carried out by means of a generalized variogram to extract vocal dysperiocities. A segmental signal-to-dysperiodicity ratio (SDRSEG) is used to predict the perceived degree of hoarseness. The corpora comprise four French sentences as well as vowels [a] produced by 22 male and female normophonic and dysphonic speakers. It is shown that the average SDRSEG correlates better with perceptual scores of hoarseness than the global signal-to-dysperiodicity ratio (SDR). The perceptual scores are based on comparative judgments by six listeners of pairs of speech samples. In addition, multiple linear regression analysis of four statistical descriptors of SDRSEG is performed to gauge whether higher-order statistics are valid predictors of perceived hoarseness.

1. INTRODUCTION

The presentation concerns the assessment of disordered voices. Acoustic feature-based assessment methods are popular in a clinical framework because they are noninvasive. Acoustic assessment enables clinicians to monitor the progress of patients and document quantitatively the perceived degree of hoarseness.

Several acoustic features have been used to characterise the speech of dysphonic speakers. Most of these features reflect the deviation of the speech signal from perfect periodicity. In the framework of disordered speech analysis, noise refers to additive noise owing to turbulence and modulation noise, including cycle-to-cycle variations of the cycle duration (jitter) and amplitude (shimmer) owing to external perturbations, as well as intrinsically irregular dynamics of the vocal folds [1]. The acoustic marker used hereafter is called signal-to-dysperiodicity ratio (SDR) [2].

Most techniques that estimate vocal dysperiodicities deal with steady fragments extracted from sustained vowels. Due to the assumptions of local stationarity and periodicity, these techniques lack robustness and accuracy when applied to connected speech produced by moderately or severely hoarse speakers.

Several authors have reported the need for extracting acoustic features from connected speech to characterise dysphonic speakers adequately [3, 4]. Indeed, in contrast to sustained vowels, connected speech includes phonetic segment onsets and offsets, as well as voiced and unvoiced intervals. Connected speech therefore enables indirectly observing the transient behavior of vocal fold vibration. Up to date, the number of studies devoted to the extraction of vocal dysperiodicities from connected speech remains small. An overview of published studies is given in [2].

In [4], a linear predictive model-based approach is proposed for analyzing vocal dysperiodicities in connected speech. Two analysis stages are used. The first is comprised of a conventional single-step linear predictive model of the speech signal and the second of a multistep predictive model of the residue that is the output of the first stage. The modeling error of the multistep predictor is used as a cue of vocal dysperiodicity.

Earlier studies have shown that a single-step linear predictive model is not, under all circumstances, appropriate for clinical analyses of speech [5]. Also, results of intelligibility tests indicate that the magnitude of the single-step linear predictive error increases with nasality or fundamental frequency, which are unrelated to vocal dysperiodicity [6].

In [2], a bi-directional three-coefficient multistep predictive model is therefore used as an alternative to the combined unidirectional single-step and multistep prediction models in [4]. The multistep linear predictive model is applied to the speech signal directly. Indeed, the bidirectional analysis avoids comparing speech fragments across phonetic boundaries, because the minimum of the left-right and right-left multistep errors is kept as a cue of vocal dysperiodicity.

Multistep linear predictive modeling exploits the local periodicity of voiced speech sounds. However, because the weights that are involved in the prediction are not constrained to be positive, multistep linear predictive analysis may invert the sign of lagged signal fragments, which is inconsistent with the definition of signal periodicity. To overcome these limitations, a generalized variogram has been proposed to estimate speech signal dysperiodicities [7].

The aim of this presentation is to examine the correlation of four statistical descriptors of the segmental signal-to-dysperiodicity ratio (SDRSEG) with the perceived degree of hoarseness. The perceptual hoarseness scores have been obtained by comparative judgments of pairs of speech samples [8]. Results show that the segmental SDRs correlate better with perceived hoarseness than the global SDRs.

The presentation is organized as follows. In Section 2, the corpus used in the experiment and the analysis methods are described. Experimental results are presented and discussed in Section 3. The conclusion is given in Section 4.

2. METHOD

2.1. Corpus

Speech data comprise sustained vowels [a], including onsets and offsets, and four French sentences produced by 22 normophonic or dysphonic speakers (10 male and 12 female speakers). The corpus

includes 20 adults (from 20 to 79 years), one boy aged 14 and one girl aged 10. Five speakers are normophonic, the others are dysphonic. Normophonic speakers' voices are modal while pathological voices range from mildly deviant to very deviant.

The sentences are the following: "Le garde a endigué l'abbé", "Bob m'avait guidé vers les digues", "Une poule a picoré ton cake" and "Ta tante a appâté une carpe". Hereafter, they are referred to as S1, S2, S3 and S4, respectively. They have the same grammatical structure, the same number of syllables and roughly the same number of resonants and plosives. Sentences S1 and S2 are voiced by default, whereas S3 and S4 include voiced and unvoiced phonetic segments.

Speech signals have been recorded at a sampling frequency of 48 kHz. The recordings were made in an isolated booth by means of a digital audio tape recorder (Sony TCD D8) and a head-mounted microphone (AKG C41WL) at the laryngology department of a university hospital in Brussels, Belgium. The recordings have been transferred from the DAT recorder to computer hard disk via a digital-to-digital interface. Silent intervals before and after each recording have been removed. Later, the recordings have been upsampled by a factor of 4 to reduce quantization noise. Indeed, simulations by means of synthetic vowels have shown that quantization noise is considerably reduced when the signal is upsampled by a factor of 4.

2.2. Perceptual assessment

The relevance of numerical cues of vocal dysperiodicity may be evaluated by their ability to predict subjective scores of hoarseness, which are based on listener perception of speech. In this presentation, a perceptual rating that is founded on comparative judgments of pairs of speech samples is used to determine the degree of hoarseness of the corpora comprised of sustained vowels [a], including onsets and offsets, and French sentences S1-S4 [8].

In the framework of a listening session, pairs of stimuli have been presented randomly to a listener who has been asked to designate the most hoarse sample of the pair. Then, the total score of the sample labeled as the most hoarse is increased by one. If both items of the pair are judged to be equally hoarse, the score of both is increased by 0.5. The experience is repeated until all possible sample pairs have been presented. At the end of the session, samples are assigned scores on the base of all possible pair comparisons. Indeed, when a sample is judged to be the most hoarse at each comparison, it cumulates the highest score at the end of a listening session. The least hoarse voice, on the contrary, is assigned a small score.

The group of judges has been comprised of six naive listeners (one female, five males), i.e. listeners without any training in speech therapy or laryngology. All have reported normal hearing. Their ages ranged from 24 to 57. They were asked to assess voice samples according to the overall degree of deviance of the voice. Each listening session has been devoted to a set of 22 stimuli. Subsequently, the average of the scores assigned by the six listeners has been selected as a subjective measure of hoarseness. Inter and intra-listener agreement is high and has been documented in [8].

2.3. Generalized variogram

The generalized variogram is derived from the conventional one by taking into account properties of the speech signal. For a periodic signal $x(n)$ of period T_0, one may write:

$$x(n) = x(n - kT_0), \quad k = \ldots -2, -1, 0, 1, 2, \ldots \tag{1}$$

A measure of the departure from periodicity over an interval of length N is an indication of the amount of signal irregularity. For stationary signals, the dysperiodicity energy may be estimated via the minimum of the following expression. The expression between brackets is known as the variogram.

$$\hat{\gamma} = \min_{T} \left[\sum_{n=0}^{N-1} \big(x(n) - x(n-T)\big)^2 \right], \tag{2}$$

with $-T_{\max} \le T \le -T_{\min}$ and $T_{\min} \le T \le T_{\max}$.

Index n positions speech samples within the analysis frame. Boundaries T_{min} and T_{max} are, in number of samples, the shortest and longest acceptable glottal cycle lengths. They are fixed to 2.5 *ms* and 20 *ms*, respectively (i.e. 50 Hz ≤ F0 ≤ 400 Hz). For voiced speech sounds, lag T is interpreted as a multiple of the speech cycle length. For unvoiced speech sounds, expression (2) remains mathematically meaningful but lag T is not interpreted in terms of the glottal cycle length.

Speech signals are expected to be locally stationary at best. The signal amplitude evolves from one speech frame to the next owing to onsets and offsets, segment-typical intensity, as well as accentuation and loudness. Introducing a weighting coefficient to account for these slow changes in signal amplitude, definition (1) becomes:

$$x(n) = a\, x(n - kT_0), \quad k = \ldots -2, -1, 0, 1, 2, \ldots \tag{3}$$

Accordingly, the generalized empirical variogram may be written as follows.

$$\hat{\gamma} = \min_{T} \left[\sum_{n=0}^{N-1} \big(x(n) - a\,x(n-T)\big)^2 \right]. \tag{4}$$

Weight a must be positive. It is defined so as to equalize the signal energies in the current and shifted analysis windows:

$$a = \sqrt{\frac{E}{E_T}}, \tag{5}$$

where E and E_T are the signal energies of the current and lagged frames,

$$E = \sum_{n=0}^{N-1} x^2(n), \quad E_T = \sum_{n=0}^{N-1} x^2(n-T).$$

The frame length N and frame shift length are equal to 2.5 *ms*. This choice guarantees that each signal fragment is included exactly once in the analysis. The instantaneous value of the dysperiodicity is estimated as follows.

$$e(n) = x(n) - a\, x(n - T_{\text{opt}}), \quad 0 \le n \le N-1, \tag{6}$$

where T_{opt} is equal to the lag which minimizes the generalized variogram (4) of the current frame position. Lag T_{opt} may be positive or negative.

2.4. Global and segmental dysperiodicity cues

The conventional signal-to-dysperiodicity ratio used to summarize the amount of dysperiodicity within an utterance is the global SDR defined as follows [2].

$$SDR = 10\log\left[\frac{\sum_{n=0}^{L-1} x^2(n)}{\sum_{n=0}^{L-1} e^2(n)}\right], \qquad (7)$$

where $e(n)$ is obtained according to (6) and L is the number of samples in the total analysis interval.

In [2], [4], [7] and [9], it has been shown that the global SDR correlates with scores obtained by means of perceptual rating of hoarseness. In this presentation, segmental SDR is examined as an alternative. One expects that the average SDRSEG of an utterance correlates more strongly with perceived hoarseness than the global SDR. Segmental SDR is known to be a good estimator of perceived quality in speech coding [10]. Since the segmental SDR values are log-weighted prior to averaging, the underemphasis in the global SDR of signal fragments that are weak and noisy is compensated for. As a consequence, low-noise high-amplitude speech sounds (e.g. stable fragments of vowels) do not numerically mask the contribution of noisy transients, for instance. For a given utterance, the analysis interval is divided into K frames of length M and the segmental SDR of each frame k is computed as follows.

$$SDRSEG(k) = 10\log\frac{\sum_{n=Mk}^{Mk+M-1} x^2(n)}{\sum_{n=Mk}^{Mk+M-1} e^2(n)}, \quad k = 0, \cdots, K-1. \qquad (8)$$

Later, multivariate analysis is carried out of four statistical descriptors of the SDRSEG and used to predict the hoarseness scores. The descriptors are the mean, variance, skewness and kurtosis [11]. In practice, they are replaced by their estimators.

Skewness and kurtosis characterise possible deviations from Gaussianity. A nonzero skewness indicates an asymetric distribution, while the kurtosis indicates whether a distribution goes asymptotically to zero more or less rapidly than a Gaussian.

3. RESULTS AND DISCUSSION

Scores of perceived hoarseness have been determined by five naïve listeners. Score averages depend slightly on the utterance type. They range from 2.2 to 7.5 for normophonic speakers and between 3.2 and 20.4 for dysphonic speakers, on a scale from 0 to 21. The effect of segment length on the strength of the correlation with the degree of perceived hoarseness has been investigated for different segment sizes (Table 1). The correlation depends sligthly on the segment length and stabilizes at 5 *ms*. The segment length has been set to this value accordingly. The averages of the SDRSEG range from 17.8 dB to 23.8 dB for normophonic speakers and from 5.5 dB to 22.3 dB for dysphonic speakers.

Pearson's product moment correlation between the degree of hoarseness and the global and average segmental SDRs of the signals corresponding to vowel [a] and sentences S1-S4 is given in Table 2. The null hypothesis ($\rho_P = 0$) has been rejected for all table entries (one-tailed test, $\rho_{crit} = 0.36$, $p < 0.05$). Inspection suggests that for sentences S1 to S3 the average segmental SDR gives rise to stronger correlations than the global SDR. The average of SDRSEG versus the corresponding degree of hoarseness is shown for sentence S1 in Fig. 1.

	[a]	S1	S2	S3	S4
30 *ms*	– 0.71	– 0.84	– 0.80	– 0.78	– 0.65
20 *ms*	– 0.71	– 0.84	– 0.80	– 0.79	– 0.67
10 *ms*	– 0.71	– 0.85	– 0.81	– 0.81	– 0.68
5 *ms*	– 0.70	– 0.86	– 0.81	– 0.81	– 0.70
2.5 *ms*	– 0.70	– 0.86	– 0.81	– 0.82	– 0.70

Table 1. Pearson's product moment correlation of average segmental SDR values with average hoarseness scores for sustained vowel [a] and sentences S1 to S4. The left column reports the segment length.

	Global SDR	Average segmental SDR
[a]	– 0.73	– 0.70
S1	– 0.72	– 0.86
S2	– 0.72	– 0.81
S3	– 0.70	– 0.81
S4	– 0.69	– 0.70

Table 2. Pearson's product moment correlation of global and average segmental SDR values with average hoarseness scores of sustained vowel [a] and sentences S1 to S4.

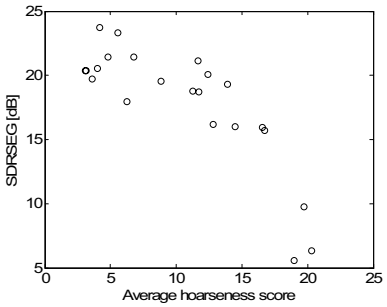
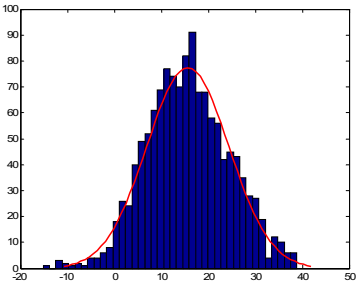


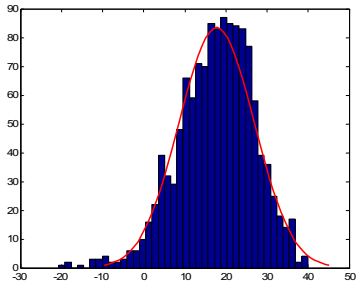
Fig. 1. Average SDRSEG vs average hoarseness scores associated with sentence S1 (22 speakers) for a segment length of 5 ms.

The lack of increase of the correlation in the case of vowel [a] is expected. Disregarding onset and offset, vocal dysperiodicities are equally distributed in time in sustained speech sounds. The lack of improvement in the case of sentence S4 is unexplained at this stage. A possible conjecture is the following. The listeners reported sentence S4 as difficult to assess because the voiced intervals appeared to be very short. Hoarse short voiced intervals would explain the equivalence of the global and segmental SDRs, because the efficacy of the segmental SDR rests on the deemphasis of high-amplitude low-noise speech fragments, which appear to be lacking in S4.

The histograms of segmental SDRs of two samples of sentence S1 are displayed in Fig. 2. Fig. 2 (a) shows the histogram corresponding to a voice that has been assigned a hoarseness score of 11.8 and Fig. 2 (b) corresponds to a voice with a hoarseness score of 3.6. The corresponding average SDRSEG values are 16 dB and 19 dB, respectively. The corresponding Gaussian curves are superimposed on the histograms. The histogram in Fig. 2 (a) appears to be closer to a Gaussian than the one in Fig. 2 (b). This is confirmed by the skewness values that are equal to 0.07 and – 0.8, respectively.



(a)



(b)

Fig. 2. Histograms of the SDRSEGs of two samples of sentence S1. (a) Hoarseness score = 11.8, average SDRSEG = 16 dB, skewness = 0.07. (b) Hoarseness score = 3.6, average SDRSEG = 19 dB, skewness = – 0.8.

Correlation analysis shows that skewness values and hoarseness scores covary. Pearson's rank correlation coefficient is equal to 0.62, which is statistically significant (one-tailed test, $\rho_{crit} = 0.36$, $p < 0.05$). This significant correlation has prompted us to investigate higher-order statistics as predictors of hoarseness scores.

To examine the relation between the hoarseness scores and the average, variance, skewness and kurtosis of the segmental SDRs, a multiple linear regression analysis has been carried out. To avoid overfitting, data of vowel [a] and sentences S1-S4 have been pooled. The standardized regression coefficients a_i, which are interpreted as factor weights of the corresponding predictor variables, and the multiple correlation coefficient are listed in Table 3. The values of the standardized regression coefficients suggest that the average segmental SDR is the most effectual predictor of hoarseness. This is confirmed by the last line in Table 3, which displays the correlations ρ_i of the individual descriptors with the hoarseness scores. Correlation ρ_1 of the average segmental SDR and multiple correlation coefficient R are quasi-identical.

a_1	a_2	a_3	a_4	R
– 0.63	– 0.16	0.10	– 0.07	0.76
ρ_1	ρ_2	ρ_3	ρ_4	—
– 0.74	– 0.31	0.37	– 0.19	—

Table 3. Standardized regression coefficients obtained by multiple linear regression analysis of the average hoarseness scores and the statistics of the SDRSEG up to order four. The multiple correlation coefficient is given to the right. The last line gives the correlation of the individual statistics with the hoarseness scores.

4. CONCLUSION

In this presentation, segmental signal-to-dysperiodicity ratios have been used to estimate vocal dyperiodicities in disordered connected speech. Experimental results show that the average segmental SDR correlates more strongly than the global SDR with the perceptual assessment of the degree of hoarseness. The ability of the statistics up to order four of the segmental SDR to predict hoarseness scores has been examined by means of a multiple linear regression analysis. The average appears to be the most effectual predictor of perceived hoarseness.

5. REFERENCES

[1] 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.

[2] F. Bettens, F. Grenez and J. Schoentgen, "Estimation of vocal dysperiodicities in connected speech by means of distant-sample bi-directional linear predictive analysis", *J. Acoust. Soc. Am,*vol. 117, no 1, pp. 328-337, 2005.

[3] F. Parsa and D. G. Jamieson., "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech ", *J. Speech, Language, and Hearing Research* , vol. 44, pp. 327-339, 2001.

[4] 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.

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

[6] M. Kahn and P. Garst, "The effects of five voice characteristics on LPC quality", *ICASSP*1983, Boston, pp. 531-534, 1983.

[7] A. Kacha, F. Grenez and J. Schoentgen, "Dysphonic speech analysis using generalized variogram", *ICASSP*2005, pp. 917-920, Philadelphia, March 2005.

[8] A. Kacha, F. Grenez and J. Schoentgen, "Voice quality assessment by means of comparative judgments of speech tokens", *Int. Conf. Spoken Lang. Process.*, Lisboa, Portugal, pp. 1733-1736, September 2005.

[9] A. Kacha, F. Grenez and J. Schoentgen, "Generalized variogram analysis of vocal dysperiodicities in connected speech", in: *Proc. of MAVEBA'*2005, Florence, Italy, pp. 155-158, October 2005.

[10] S. R. Quackenbush, T. P. Barnwell III and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall Englewood Cliffs, 1988.

[11] C. L. Nikias and A. P. Petropulu, *Higher-order spectral analysis*, Prentice Hall, Englewood Cliffs, 1993.