Multivariate Analysis of Frame-Based Acoustic Cues of Dysperiodicities in Connected Speech

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

[(*)] Department Signals and Waves, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium
[(**)] National Fund for Scientific Research, Belgium
akacha@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be

Abstract

Generalized variogram is used to extract vocal dysperiocities in disordered speech produced by dysphonic speakers. Both signal and dysperiodicity are passed through a filter bank and a segmental signal-to-dysperiodicity ratio is defined in each frequency band. Multivariate analysis is carried out to summarize the degree of perceived hoarseness. The predictor variables are the segmental signal-to-dysperiodicity ratios in the different bands. It is shown that high correlations are achieved by linear regression analysis of the segmental signal-to-dysperiodiciy ratios in different non-overlapping frequency bands.

Index Terms : disordered speech, multivariate analysis, generalized variogram, dysperiodicity estimation.

1. Introduction

One of the manifestations of voice disorders is the lack of periodicity in voiced speech produced by dysphonic speakers. Dysperiodicities may be caused by additive noise owing to turbulence and modulation noise owing to external perturbations of the glottal excitation signal, as well as dysperiodicities due to intrinsically irregular dynamics of the vocal folds. Several acoustic features used to assess voice disorders reflect the deviation of the speech waveform from the perfect periodicity. For instance, jitter and shimmer are frequently used to measure perturbations produced by the variations in the fundamental period and amplitude, respectively. One acoustic marker of hoarseness is the so called signal-to-dysperiodicity ratio (SDR) [1].

Most techniques for estimating vocal dysperiodicities have been applied to steady fragments extracted from sustained vowels. The widespread use of sustained vowels is due to the technical feasibility of the analysis rather than clinical relevance [2].

As the evaluation of the voice quality is usually based on the perception of continuous speech, it is expected that acoustic features extracted from continuous speech correlate better with the perceptual assessment of the voice quality. Indeed, connected speech contains the dynamic characteristics of voice source and vocal tract such as voice onset and offset and variation in fundamental frequency and amplitude [2].

In [3], a generalized variogram has been proposed to estimate vocal dysperiodicities in connected speech as an alternative to the multistep linear predictive modeling developed in [1] and [4].

The conventional marker used to summarize vocal noise (dysperiodicities) within an utterance is the global signal-to-dysperiodicity ratio (SDR). Global signal-to-dysperiodicity ratio is mostly contributed by the vocalic segments in connected speech [1]. In [5], it has been found that segmental signal-to-dysperiodicity ratio (SDRSEG) outperforms global signal-to-dysperiodicity ratio in terms of correlation with perceived hoarseness. Segmental SDR is calculated on a frame basis and the average is carried out so as to favor short, weak and noisy frames. It, therefore, reports more accurately what is typical of connected speech, i.e., transients and unsteady speech fragments that are expected to report on vocal cyclicity in a different manner than pseudo-steady vocalic phonetic segments.

In [6], a multiparametric method has been used to analyze sustained vowels [a]. It has been shown that a nonlinear combination of six acoustic parameters resulted in 86 % concordance with perceptual evaluation.

The aim of this presentation is to propose a multivariate analysis to summarize the degree of perceived hoarseness. Both the signal and dysperiodicity estimated by means of the generalized variogram are passed through a filter bank and a segmental signal-to-dysperiodicity ratio is defined in each frequency band. The predictor variables involved in the multivariate analysis are the segmental signal-to-dysperiodicity ratios in the different frequency bands. It is shown that high correlations with perceived hoarseness are achieved by linear regression analysis of the segmental signal-to-dysperiodiciy ratios in different non-overlapping frequency bands.

2. Methods

2.1. Corpora

A first corpus comprises 1-second stationary fragments of vowel [a] sustained by 38 adult normophonic speakers (22 male and 16 female) and 51 adult dysphonic speakers (19 male and 32 female). The patients have been diagnosed on the base of a clinical examination at the ENT department of the Erasme University Hospital in Brussels, Belgium. The recordings have been made by means of a PCM Sony audio-processor at a sampling frequency of 48 kHz and stored on video tape. Subsequently, they have been low-pass filtered, resampled at 20 kHz and stored on computer disk for further processing.

A second corpus comprises 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. The dysphonic speakers were patients of the Laryngology Department of the Erasme University Hospital in Brussels, Belgium.

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 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). 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.

2.2. Perceptual rating

For the first corpus comprising stationary fragments of sustained [a], the degree of hoarseness has been determined by five judges on the base of a visual inspection of the spectrograms. Each judge has assigned a degree between 0 and 4. Consequently, the overall degrees of perceived hoarseness have been comprised between 0 and 20. Spectrograms that have been considered to be noisy or clean have thus been assigned high or low degrees of hoarseness respectively. The selection of the judges, as well as the intra-judge and inter-judge agreement have been presented in [7].

For the second corpus comprising vowels [a] and sentences S1 to S4, A perceptual rating that is founded on comparative judgments of pairs of speech tokens has been used to determine the degree of hoarseness [8]. The perceptual assessment exploits the ability of listeners to compare two stimuli in terms of grade, i.e., perceived overall degree of deviance of the voice. The aim is to hierarchize a set of recordings from the least to the most anomalous by means of comparative judgments of all possible token pairs within the set. The procedure is the following.

1. The list of all possible different pairs of items is formed. An item is a recording belonging to a set of identical stimuli.
2. All scores are initialized to zero.
3. A randomly selected pair of speech recordings is presented to the listener, who is asked to point out the recording with the highest perceived hoarseness. The listener has also the option to label both recordings of a pair as equally hoarse.
4. The total score of the recording 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 recordings is increased by 0.5.
5. Steps 3 and 4 are repeated until all possible pairs that belong to a same session have been presented.
6. The speech tokens are ranked on the base of their total score, i.e., the speech token that has been labeled the most often as the most abnormal of the pairs is assigned the highest rank and the speech token that has been the most often labeled as the least abnormal of the pairs is assigned the lowest rank.

All the pairs of stimuli produced by 22 speakers have been compared, i.e a total of 231 pairs. Speech tokens have been presented via a digital-to-analog audio interface (Digidesign Mbox) and dynamic stereo headphones (Sony MDR-7506). Loudness has been fixed at a comfortable level by the listener.

The group of judges has been comprised of six naïve listeners, i.e. listeners without training in speech therapy or laryngology. All reported normal hearing. Their ages ranged from 24 to 57. One listening session was devoted to a set of 22 stimuli. The total number of sessions has been equal to 6 listeners x 5 stimuli = 30. The same experiment has been repeated by four listeners after a period of a day at least to gauge intra-judge reliability. The total number of retest sessions has therefore been equal to 4x5=20.

The average of the scores assigned by the six listeners has been selected as a subjective measure of perceived hoarseness. The intra-judge and inter-judge agreement has been presented in [8]. Correlation analysis showed high inter and intra-listener agreement.

2.3. Generalized variogram

The generalized variogram is derived from the conventional one by taking into account properties of the speech signal [3]. For a periodic signal $x(n)$ of period T_0, one may write:

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

A measure of the departure from periodicity over an interval of length N is an indication of the amount of signal irregularity.

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:

$$x(n) = a\,x(n - kT_0), \quad k = \ldots -2, -1, 0, 1, 2, \ldots \qquad (2)$$

The dysperiodicity energy may be estimated via the minimum of the following expression. The expression between brackets is the empirical generalized variogram.

$$\hat{\gamma} = \min_{T} \left[\sum_{n=0}^{N-1} \left(x(n) - a\,x(n-T) \right)^2 \right]. \qquad (3)$$

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

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

$$a = \sqrt{E/E_T}, \qquad (4)$$

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

$$E = \sum_{n=0}^{N-1} x^2(n), \; E_T = \sum_{n=0}^{N-1} x^2(n-T). \qquad (5)$$

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). 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.

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

where T_{opt} is equal to the lag which minimizes the generalized variogram (3) of the current frame position. Lag T_{opt} may be positive or negative.

2.3. Multivariate analysis of frame-based acoustic cues

The conventional acoustic marker commonly used to summarize the amount of dysperiodicity within an utterance is the global signal-to-dysperiodicity ratio defined as follows [1]:

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

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

In [5], a short-time calculation of the SDR measure has been proposed as an alternative and proven to outperform the global SDR in terms of correlation with scores of perceived hoarseness. In this presentation, multivariate analysis carried out on segmental signal-to-dysperiodicity ratios is used to summarize the amount of perceived hoarseness. Both the signal and dysperiodicity are filtered in a B-band filterbank and the segmental SDR is computed in each band. Segmental SDR in each band is obtained by reformulating locally the conventional global SDR, i.e., the marker is repeatedly computed over short segments of the analysis interval and the average is used as an acoustic cue of perceived quality.
Denote by SDR_j the segmental signal-to-dysperiodicity ratio in the band j [9],

$$SDR_j = \frac{10}{K}\sum_{k=0}^{K-1} \log \frac{\sum_{n=Mk}^{Mk+M-1} x_j^2(n)}{\sum_{n=Mk}^{Mk+M-1} e_j^2(n)},\ j = 1,\cdots,B \qquad (8)$$

where $x_j(n)$ and $e_j(n)$ are the filtered signal and instantaneous dysperiodicity in the band j. M and K are the frame length and frame number in the overall analysis interval, respectively.
Segmental SDRs from the different bands are used as variable predictors of scores of perceived hoarseness

$$Score = c + \sum_{j=1}^{B} b_j SDR_j \qquad (9)$$

The parameters c and b_j are the regression coefficients obtained by minimizing the mean square error in the prediction of the degree of perceived hoarseness as a linear combination of the segmental SDRs in the different bands.

The local formulation is expected to be more adequate than the global formulation in terms of the correlation of the acoustic cue with perceived hoarseness. The reason is that the segmental SDR values are log-weighted prior to averaging, which compensates for the underemphasis in the global SDR of signal fragments that are weak and noisy. 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.

A linear phase FIR filterbank has been used to decompose the signal, as well as the dysperiodicity into three non-overlapping frequency bands B_1 to B_3, and subsequently, the acoustic measures are computed according to (8). The use of three bands avoids overfitting. These frequency bands have been selected to cover the range of perceived tones and noises. Their ranges are (0 – 2500 Hz), (2500 Hz – 5000 Hz) and beyond 5000 Hz.

3. Results

In this Section, the performance of the multivariate analysis in predicting scores of hoarseness is investigated and compared to that of the segmental SDR (univariate analysis) used in [5].

Generalized variogram analysis has been carried out on speech signals corresponding to sustained vowels [a] produced by 89 speakers as well as vowels [a], including offset an offset, and four sentences S1 to S4 produced by 22 speakers. Table 1 gives Pearson product moment correlations of segmental SDRs with average scores of hoarseness for both corpora. The effect of frame length on the strength of the correlation with the degree of perceived hoarseness has been investigated for different frame sizes. It has been found that the correlation depends slightly on the segment length and stabilizes at 5 ms. The frame length has been set to this value accordingly.

The speech signals and dysperiodicities have been passed through the three-channel filterbank and multivariate analysis has been carried out on segmental SDRs from different frequency bands. Table 2 displays the results of the linear regression analysis carried out on segmental SDRs from the different frequency bands. The Table reports standardized regression coefficients β_i, which are interpreted as factor weights of the corresponding predictor variables and the multiple correlation coefficients for both corpora. Multiple correlation coefficients are statistically significant for sustained vowels of the first corpus ($R_{\text{crit}} = 0.30$, $p < 0.05$) and for sustained vowels [a] as well as for sentences S1 to S4 of the second corpus ($R_{\text{crit}} = 0.59$, $p < 0.05$). Data of vowels [a] and sentences S1 to S4 of the second corpus have been pooled to form a single sequence and linear regression analysis has been carried out on the pooled data. The results are listed in the last row of Table 2. The Multiple correlation coefficient is statistically significant ($R_{\text{crit}} = 0.27$, $p < 0.05$). The adjusted R^2 provides a more conservative estimate of the percentage of variance in the criterion variable that can be attributed to the combined predictor variables.

To compare the contribution of the different predictor variables, multiple correlations obtained from linear regression analysis carried out on segmental SDRs from one, two and three bands are displayed in Table 3.

4. Discussion

Correlation analyses show that linear regression analysis of segmental SDRs from different frequency bands results in an improvement of the performance in terms of correlation with scores of perceived hoarseness over segmental SDR. Inspection of Tables 2 and 3, shows that multivariate analysis gives rise to stronger correlation for sustained vowels [a] of the first and second corpora as well as for sentence S4. The

improvement of the correlation for the first corpus is attributed to the multivariate analysis approach rather than to the use of the perceptual rating based on visual inspection of spectrograms. Indeed, the correlation is stronger than that obtained in [7] by carrying out multivariate analysis on flat spectra.

The high multiple correlations suggests that in their rating, listeners are influenced by short noisy fragments, as well as by the acoustically prominent vocalic segments. From Table 3, one sees that the significant contribution in the prediction of hoarseness scores is due to the segmental SDR from the band B_1. Segmental SDRs (univariate analysis) and segmental SDRs in the frequency band B1 are similarly correlated with perceived hoarseness. Segmental SDRs estimated from the bands B_2 and B_3 give rise to an increase of the correlation that depends on the utterance. For sustained vowel [a] of the first corpus, the correlation is increased from 0.84 to 0.90 by using segmental SDR from the frequency band B2. This increase corresponds to 10 percent of the variance in the hoarseness scores. The segmental SDR obtained from band B_2 makes a significant contribution for sustained vowels [a] resulting in an increase of the multiple correlation from 0.73 to 0.77. This increase of the correlation due to the inclusion of segmental SDR in the band B_2 corresponds to 6 percent of the variance in the hoarseness scores. However, the use of segmental SDR from the frequency band B3 does not improve the correlation for vowels [a] of the both corpora. For sentence S4, the use of segmental SDR estimated from the band B_3 results in a significant contribution. This contribution accounts for an additional 13.3 percent of the variance in the predicted variable (hoarseness scores).

Table 1: Correlations of segmental SDRs with average scores of perceived hoarseness for sustained vowels [a] of the first corpus (89 speakers) and for sustained vowels [a] as well as sentences S1 to S4 of the second corpus (22 speakers).

[a] (89)	[a] (22)	S1	S2	S3	S4
–0.85	–0.70	–0.86	–0.81	–0.81	–0.70

Table 2: Results from linear regression analysis carried out on segmental signal-to-dysperiodicity ratios from three non-overlapping frequency bands. The dependent variable is the score of perceived hoarseness.

	β_1	β_2	β_3	R	R^2	Adj. R^2
[a] (89)	–0.48	–0.55	0.09	0.90	0.81	0.80
[a] (22)	–0.52	–0.37	0.14	0.78	0.61	0.55
S1	–0.70	–0.20	0.25	0.87	0.76	0.72
S2	–0.77	–0.10	0.12	0.82	0.68	0.62
S3	–0.66	–0.32	0.18	0.83	0.69	0.64
S4	–0.40	–0.44	0.48	0.78	0.61	0.55
Pooled data	–0.56	–0.34	0.30	0.78	0.61	0.60

5. Conclusion

In this presentation, the performance of the linear combination of segmental SDRs estimated in non-overlapping frequency in terms of correlation with scores of perceived hoarseness has been examined and compared to that of the segmental SDR. Experimental results show that linear regression analysis results in a higher correlation with scores of perceived hoarseness for sustained vowels as well as for connected speech. It is expected that the proposed measure correlates with other acoustic measures such as jitter, shimmer and harmonics-to-noise ratio.

In this study, it has been assumed that dysperiodicities are caused by vocal disorders so that background noise can affect the value of the acoustic measure. Improvement of the algorithm by incorporating an appropriate model to account for background noise may be considered in the future work.

Table 3: Results obtained from linear regression analysis carried out on segmental SDRs from one, two and three bands. Multiple correlations of the predicted hoarseness scores with perceptual ratings.

	B1	B1, B2	B1, B2, B3
[a] (89)	0.84	0.90	0.90
[a] (22)	0.73	0.77	0.78
S1	0.85	0.85	0.87
S2	0.82	0.82	0.82
S3	0.81	0.82	0.83
S4	0.68	0.69	0.78
Pooled data	0.74	0.75	0.78

6. References

[1] Bettens, F., Grenez, F., and Schoentgen, J., "Estimation of vocal dysperiodicities in connected speech by means of distant-sample bi-directional linear predictive analysis", *J. Acoust. Soc. Amer.,* 117(1), pp. 328-337, 2005.

[2] 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.

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

[4] Qi, Y., Hillman, R. -E., and Milstein, C., "The estimation of signal-to-noise ratio in continuous speech of disordered voices", *J. Acoust. Soc. Amer.,* 105(4), pp. 2532-2535, 1999.

[5] Kacha, A., Grenez, F., and Schoentgen, J., "Frame-based acoustic cues of vocal dysperiodicity in connected speech", in: *Proc. ICASSP'*2006, pp. 385-388, Toulouse, May 2006.

[6] Yu, P., Quaknine, M., Revis, J., and Giovanni, A., "Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements", *J. Voice*, 15(4), pp. 529-542, 2001.

[7] Schoentgen, J., Bensaid, M., and Bucella, F., "Multivariate statistical analysis of flat vowel spectra with a view to characterizing dysphonic voices" *J. Speech, Lang., Hear. Res*., 43(6), pp. 1493-1508.

[8] 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.

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