Multi-band Segmental Signal-to-dysperiodicity Ratios in Connected Speech Produced by Normophonic and Dysphonic Speakers

A. Alpan[1], F. Grenez[1], M. Remacle[2], J. Schoentgen[1,3]

[1] Laboratories of Images, Signals and Telecommunications, Université Libre de Bruxelles, Brussels, Belgium

[2] Department of Otorhinolaryngology and Head and Neck Surgery, University Hospital of Louvain at Mont-Godinne, Yvoir, Belgium

[3] National Fund for Scientific Research, Belgium

aalpan@ulb.ac.be, fgrenez@ulb.ac.be, remacle@orlo.ucl.ac.be, jschoent@ulb.ac.be

Abstract

The objective is to analyze vocal dysperiodicities in connected speech produced by dysphonic speakers. The analysis involves a variogram-based method that enables tracking instantaneous vocal dysperiodicities. The dysperiodicity trace is summarized by means of the signal-to-dysperiodicity ratio, which has been shown to correlate strongly with the perceived degree of hoarseness of the speaker. Previously, this method has been evaluated on small corpora only. In this article analyses have been carried out on a corpus comprising over 700 speaker, which is split into normophonic and pathological speakers. First, statistically significant differences have been found between the averages of the full-band signal-to-dysperiodicity ratios of the normal and disordered utterances. Second, multi-band signal-to-dysperiodicity ratios have been submitted to principal component analysis. Results show that the first two principal components are interpretable in terms of the degree of dysphonia and the spectral slope respectively. The clinical relevance of the principal components has been confirmed by linear discriminant analysis.

1 Introduction

Acoustic analysis of speech is non-invasive and enables clinicians to monitor and express numerically the degree of hoarseness of a speaker's voice.

Many voice disorders cause voiced speech to deviate from strict periodicity. Dysperiodicities may be caused by additive noise owing to turbulent airflow and modulation noise owing to extrinsic perturbations of the glottal excitation signal. Dysperiodicities may also be due to an intrinsically irregular dynamics of the vocal folds or involuntary transients between dynamic regimes.

Many acoustic features that have been used to assess vocal function reflect the deviation of the speech waveform from perfect periodicity. Most of them have been obtained for steady fragments of sustained vowels, owing to technical feasibility rather than clinical relevance [1]. However, clinicians consider connected speech to be more informative than sustained vowels.

The generalized variogram method enables tracking cycle-to-cycle dysperiodicities (whatever their cause) in any speech sound produced by any speaker, because it is not based on the assumptions that the signal is locally periodic or that the average period length can be known in advance [2]. The signal-to-dysperiodicity ratio (SDR) that summarizes the dysperiodicities has been shown to correlate strongly with the degree of perceived hoarseness.

Previously, the variogram-based method has been tested on small corpora. In this presentation, the signal-to-dysperiodicity ratios are obtained for a corpus of sustained vowels and connected speech fragments produced by over 700 speakers. This enables performing multi-band frequency analysis without risking over-fitting. The objectives of the

experiments involve the following: a) test whether the averages of the full-band signal-to-dysperiodicity ratios are significantly different for normophonic and pathological speakers; b) investigate, via a principal component analysis, signal-to-dysperiodicity ratios obtained in different frequency bands. The clinical relevance of the principal components has been tested via a linear discriminant analysis of speech tokens known to be "normal" or "pathological".

2 Methods

2.1 Extraction of vocal dysperiodicities

The variogram method is based on the observation that when one reports in a 2-dimensional graph samples of a noise-free periodic signal on the horizontal axis and samples that are identically positioned in an adjacent period on the vertical axis, then all sample pairs (x,y) are located on the bisector of the graph. In a noisy signal, pairs (x,y) remain in the vicinity of the bisector. The cumulated distance between pairs and bisector over an analysis frame is a measure of the total signal noise in that frame and the individual distances between each pair and the bisector are sample-by-sample estimates of the noise (whatever its cause).

In practice, a sliding rectangular analysis frame of 2.5 ms is used and auxiliary frames are time-shifted to the left and right to minimize the cumulated distance of all inter-frame sample pairs. This inter-frame distance calculated for all inter-frame shifts is known as the variogram, from which the method takes its name. The positioning of analysis frames to the left and right of the main analysis frame avoids comparing signal fragments that do not belong to the same phonetic segment because the minimum of the left and right distances is retained as a measure of vocal noise [2].

Before the calculation of the individual and cumulated distances, the within-frame signal fragments are energy-normalized. Energy-normalization enables compensating for slow amplitude variations.

To obtain vocal dysperiodicity estimates for a complete signal, the main frame is shifted without overlap or gap and the analysis is repeated as often as necessary.

2.2 Segmental signal-to-dysperiodicity ratio

The vocal noise is summarized by means of segmental signal-to-dysperiodicity ratios. Speech signal *x(n)* as well as the corresponding dysperiodicity trace *e(n)* are divided into intervals of length *Ls* equal to 5 ms [2]. Then, a local signal-to-dysperiodicity ratio (1) is computed for each interval.

$$SDR_{loc} = 10\log\frac{\sum_{n=0}^{Ls-1} x^2(n)}{\sum_{n=0}^{Ls-1} e^2(n)} \quad (1)$$

The segmental signal-to-dysperiodicity ratio *SDRSEG* is obtained by averaging the SDR_{loc}s over all intervals.

2.3 Multi-band analyses

For each utterance, the speech signal as well as the corresponding dysperiodicity trace are filtered by means of four mel-spaced linear-phase filters and segmental signal-to-dysperiodicity ratios (1) are computed for each band. The ranges of the four mel bands (B1 – B4) are (0 – 800 mel), (800 – 1600 mel), (1600 – 2400 mel) and beyond. They correspond to the frequency bands (0 – 724 Hz), (724 – 2195 Hz), (2195 – 5188 Hz) and beyond.

2.4 Corpus

The corpus has been the Kay Elemetrics Voice Disorder Database developed by the Massachusetts Eye and Ear Infirmary Voice and Speech Labs. This corpus comprises 53 normal and over 650 disordered utterances. The acoustic tokens are sustained phonations of vowel [a] (3 - 4 s long) and the first 12 seconds of the Rainbow Passage spoken by normophonic subjects and patients with organic, neurological, traumatic, and psychogenic voice disorders at different stages (from early to fully developed). No perceptual evaluation of the tokens is available.

2.5 Statistical analyses

Firstly, hypothesis tests have been performed to check whether the averages of the full-band segmental signal-to-dysperiodicity ratios are significantly different for normophonic and dysphonic speakers. Secondly, analyses of variance have been carried out to compare the averages of the

full-band segmental signal-to-dysperiodicity ratios between different categories of pathologies.

2.6 Principal component and linear discriminant analyses

A principal component analysis has been carried out on 3 segmental signal-to-dysperiodicity ratios (for the three lowest bands) of the corpus comprising normal and disordered utterances. The *SDRSEG* cues have been z-normalized prior to analysis.

A linear discriminant analysis has been performed to assess numerically the clinical relevance of the first two principal components.

3 Results

3.1 Statistical analyses

Statistically significant differences have been observed between the averages of the full-band segmental signal-to-dysperiodicity ratios of the normal and disordered utterances for sustained [a] (two-tailed *t*-test, t=24.43, p<0.001) and the 12-second Rainbow Passage (two-tailed *t*-test, t=13.78, p<0.001). With regard to the comparison of different categories of pathologies via the analysis of variance, statistically significant differences have been observed between several pathology categories. Tukey tests have been used as post-hoc multiple comparison tests [3]. With regard to vowel [a], one observes that the vocal nodule category statistically significantly differs from the severe ventricular compression and bowing pathology categories. However, for the Rainbow Passage, Tukey tests have not detected any statistically significant differences between pathology categories.

3.2 Principal component and linear discriminant analyses

Table 1 shows, for the Rainbow Passage, the results of the principal component analysis applied to the *SDRSEG*s obtained for the first three frequency bands. Eigenvalues as well as cumulative variances are shown to the left. Coefficients of the linear combinations, which transform the (z-normalized) *SDRSEG*s into principal components, are shown to the right.

One observes that more than ninety percent of the total variance are explained by the first two principal components PC_1 and PC_2, which are interpreted as the negative of the average of the z-normalized *SDRSEG*s and the difference between the z-normalized *SDRSEG*s in bands 3 and 1, respectively.

Table 1: *Results of the principal component analysis applied to the segmental signal-to-dysperiodicity ratios (SDRSEG) obtained for the first three frequency bands of the connected speech corpus (12-second of the Rainbow Passage).*

PC	Eigen-values	Cumulative variance	Coefficients (for the different bands)		
			1	2	3
1	2.00	66.8 %	-0.56	-0.66	-0.50
2	0.79	92.9 %	-0.64	-0.05	0.77
3	0.21	100.0%	-0.53	0.75	-0.40

When one reports the second versus the first principal component in a 2D graph, normal and disordered utterances tend to cluster separately (Figure 1). A principal component analysis of *SDRSEG*s of sustained [a] give similar results.

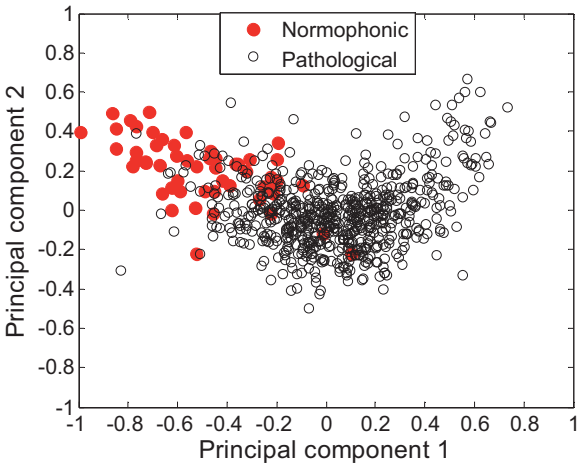

Figure 1: *Principal component representation of the SDRSEGs of the Rainbow Passage corpus.*

A linear discriminant analysis has been carried out to assess numerically the clinical relevance of the first two principal components. For the Rainbow Passage, one observes that out of 661 disordered tokens, 596 have been correctly classified as disordered and 65 have been misclassified as normal.

Similarly, out of 53 normal tokens, 46 have been correctly classified as normal and 7 have been misclassified as disordered. Thus, the overall classification accuracy is 89.9%. Here, linear discrimination analysis is carried out to confirm the relevance of the principal component analysis. One does not suggest detecting dysphonic speakers automatically.

4 Discussion and conclusion

Principal component analyses have been carried out on the multi-band *SDRSEG* cues. One observes separate clustering of disordered and clean tokens (Figure 1). The first principal component indeed corresponds to the negative of the average z-normalized signal-to-dysperiodicity ratios in the three mel-frequency bands (Table 1). Clean voices are therefore assigned to the left and severely disordered voices to the right of the horizontal axis (Figure 1).

The crescent shape of the graph can be interpreted in terms of the spectral slopes of the speech spectra. To illustrate, Figure 2 shows two vowel [a] spectra that correspond to tokens with $PC_1<0$ and $PC_2<0$ as well as $PC_1<0$ and $PC_2>0$. When comparing the spectrum in Figure 2.a ($PC_2>0$) to the one in Figure 2.b ($PC_2<0$), one sees that in spectrum Figure 2.b the harmonics decrease more rapidly with frequency than in Figure 2.a. This suggests that the second principal component depends on the spectral slope. Indeed, the spectral slopes for $PC_2<0$ are steeper (the slope is larger in absolute value) than for $PC_2>0$. This observation agrees with the interpretation of principal component 2 as a difference between the *SDRSEG*s in frequency bands 3 and 1 (Table 1).

The spectra reported in Figure 2 correspond to normal voices. Indeed, their harmonic structure is well defined and the dB level of the speech spectrum is higher than the dB level of the dysperiodicity spectrum. This agrees with the interpretation of the first principal component as an average that reports the overall degree of dysperiodicities.

In addition, an overall classification accuracy of 89.9% has been obtained via linear discriminant analysis based on the first two principal components. This suggests that principal components combine dysperiodicity cues in a clinically meaningful way.

5 Acknowledgements

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

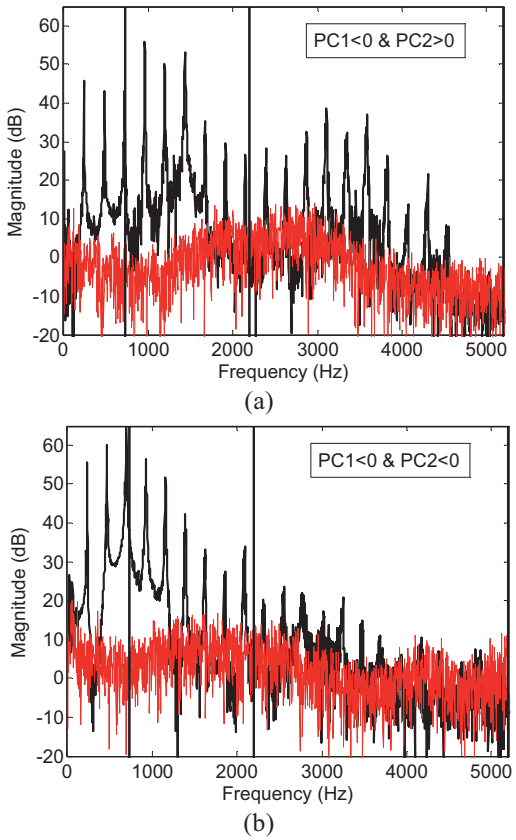

Figure 2: *Speech and dysperiodicity signals spectra. In black: Speech spectrum, in grey: dysperiodicity spectrum.*

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] D.E. Hinkle, W. Wiersma, S.G. Jurs, Applied statistics for the behavioral sciences, New York: Houghton Mifflin Company, 1998