DESCRIBING HOARSE VOICES BY MEANS OF A MULTIVARIATE ANALYSIS OF VOWEL SPECTRA

Fabrizio Bucella**, Jean Schoentgen+*, Mounir Bensaid
**Laboratory of Experimental Phonetics, Université Libre de Bruxelles, Brussels, Belgium, +National Fund for Scientific Research, Belgium, **Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture, Belgium*

ABSTRACT

The presentation concerns the description of hoarse voices by means of spectral cues obtained via a principal component analysis of contour-free spectra. Conventional spectral descriptors of hoarse voices involve the segregation of the harmonics from the rest of the spectrum. Possible problems are the following. First, the segregation of the harmonics is based on heuristic methods that make it possible to achieve an approximate separation only. Second, the spectral contour conditions the magnitudes of the harmonics and inter-harmonics and, therefore, their ratio that is involved in the spectral descriptors of hoarseness. We therefore propose describing spectra of hoarse voices by removing the spectral contour, ranking the spectral components according to their size, grouping components of similar size and performing a principal component analysis of the energies of the grouped spectral components. Results show that the first two principal components enable severely hoarse, moderately hoarse and clear voices to be distinguished.

1. INTRODUCTION

This presentation is mainly concerned with methodological issues pertaining to the characterization of hoarse voices by means of spectral descriptors. Hoarse voices have been described via spectral cues that are founded on the ratio of harmonic and inter-harmonic spectral energies. The reason is not that the harmonics stand in for the signal and the inter-harmonic for the noise. The reason is rather that with increasing noise the magnitudes of the harmonics and inter-harmonics decrease and increase respectively. Expressions involving the ratio of the inter-harmonics and harmonics therefore vary monotonically with noise. Spectral cues of hoarseness proposed in the literature differ mainly in the heuristic processing that has been used to discover the harmonics and isolate them from the rest of the spectrum [1-3].

Other approaches to the description of noise in the speech signal are the one by Michaelis et al. who defined a relative measure of the additive and modulation noise [4], and the temporal methods that rely on the assumption that aligning and adding many cycles of sustained vowels yields a noise-free cycle to which noisy cycles can be compared [5,6].

Possible problems with conventional spectral cues are the following. First, the search for and segregation of the harmonics from the rest must be performed heuristically since both the harmonics and inter-harmonics are affected by increasing noise and no statistical criteria exist that would distinguish the one from the other. In addition, for a given level of noise, the harmonics and inter-harmonics decrease and increase respectively towards the right of the frequency axis till the harmonics blend into the inter-harmonic baseline. This blending occurs gradually and not abruptly. Consequently, the decision to assign a spectral component to the harmonics or to the noise is arbitrary up to a certain point. Second, the spectral contour co-determines the magnitudes of all spectral components and therefore the spectral cues of hoarseness.

The spectral contour is the smooth line drawn through the spectral baseline. The contour differs from the spectral envelope, which is the smooth line drawn through the tops of the harmonics whose sizes depend on the amount of additive or modulation noise. The spectral contour, however, is irrelative to the description of the spectral consequences of noise. This might explain the choice of some to focus on short spectral intervals over which the spectral contour is not expected to vary much.

We here propose an alternative that is based on the ranking of the contour-less spectral components according to their size. The removal of the spectral contour frees the spectral magnitudes from the influence of the shape of the glottis signal and the vocal tract transfer function. Consequently, the spectral baseline is flattened and the sizes of the harmonics and inter-harmonics can be compared directly. The ranking therefore positions the harmonics together because they stick out above the flattened baseline. The search for a mathematical expression that varies with hoarseness and that is based on the sizes of the spectral components can then be performed by means of a multivariate analysis of the grouped magnitudes. Multi-variate data analysis indeed makes it possible to detect hoarseness-related patterns in the (grouped) spectral data, subject to the condition that hoarseness is the main source of inter-spectral variability (which is likely to be the case considering that the spectral contour has been removed).

2. METHODS

The methods section is divided into sub-sections concerning the corpora, the perceptual classification, the signal processing and the statistical analysis of the spectral data.

2.1 Corpus

The corpus was comprised of vowels [a] [i] and [u] sustained by 38 French-speaking subjects (21 males and 17 females) who did not report any laryngeal problems and 51 speakers (32 females and 19 males) who were patients in the ENT department of the University Hospital. The healthy subjects were from the University staff or were recruited by advertisement. The average ages were 46 and 49 for the healthy males and females, and 48

and 47 for the dysphonic males and females. The pathologies, diagnosed by the ENT department's doctors, fell into the following categories: edema of the vocal folds, nodules, hypotonia and asthenia, pseudo-cysts, granuloma, congestion of the vocal folds, polyps, chronic laryngitis, hyperkinesia, and paralysis.

Vowels [a] [i] [u] were repeated in that order at least three times. Each speaker was instructed to sustain the vowels at comfortable pitch and loudness levels and to avoid getting out of breath. The recordings were carried out in an isolated booth by means of a Sennheiser microphone mounted on a headset, thus keeping the distance between the mouth and the microphone constant. The signals were digitized at 48 kHz via a PCM Sony audio-processor and stored on videotape. Later, a stable portion of each vowel was chosen and redigitized at 20 kHz at the resolution of 12 bits and stored on computer disk for further processing. Before resampling, the signals were filtered by means of an anti-aliasing filter with a cut-off frequency of 10 kHz.

One vowel segment of each timbre was selected for noise characterization via the spectrum. The final corpus was therefore made up of 114 vowels produced by the healthy and 153 vowels produced by the dysphonic speakers.

2.2 Perceptual evaluation

To arrive at a referential classification pertaining to hoarseness, a panel of five judges visually classified the wideband spectrograms of vowels [a], [i] and [u] according to the scheme proposed by Yanagihara in [7]. The spectrograms were presented randomly on a computer screen. The judges responded by clicking on buttons labeled from 0 to IV, degree 0 corresponding to clear voices and degree IV to severely hoarse ones. The objective was to arrive at an independent classification to which the results obtained by multi-variate analysis could be compared and by means of which it could be interpreted. The objective was not to study the (visual) perception of hoarse voices per se.

2.3 Signal processing

The signal processing comprised the following stages. Steps (a) and (b) relate to the calculation of the magnitude spectrum. Steps (c) to (e) refer to the removal of the spectral contour by means of a wavelet transform of the squared magnitude spectrum. Wavelets are functions that oscillate about zero, but the oscillations damp down to zero so that the function is localized in frequency, time, or position (depending on the physical dimension of the independent variable [8].) Finally, steps (f) to (h) are corrective only. Their purpose was to compensate for the slight negative dip at low frequencies of the baseline of the contour-free spectrum. The dip is a consequence of the average of the contour-free spectrum being zero. The negative dip compensates for the strong (but narrow) harmonics that at low frequencies stand out above the baseline.

The signal processing steps were the following.

(a) Multiplication of a sustained vowel by a Hamming window 16384 samples long. The sampling frequency was 20 kHz.
(b) Performance of the Fast Fourier Transform and the calculation of the squared magnitude spectrum.
(c) Performance of the Wavelet Transform of the squared magnitude spectrum of length 16384.
(d) Assignment of the first 128 wavelet coefficients (constant included) to the spectral contour, and of the remaining 16254 to the contour-less spectrum.
(e) Acquisition of the contour-less squared magnitude spectrum by means of the Inverse Wavelet Transform of the 16254 wavelet coefficients.
(f) Representation of the contour-less squared magnitude spectrum in doubly logarithmic coordinates.
(g) Fitting of a straight-line to the baseline of the doubly logarithmic contour-less spectrum.
(h) Subtraction of the fitted line to obtain the corrected contour-free squared magnitude spectrum.

2.4 Statistical analysis

We performed a principal component analysis of the magnitudes of the contour-free spectral components. Principal component analysis transforms the original variables into new ones that are uncorrelated and account for decreasing proportions of the variance in the data. The new variables, the principal components, are defined as linear combinations of the original variables [9].

Ideally, the original variables are the spectral magnitudes. In practice, the number of original variables must be at most a small fraction of the number of spectra (or speakers) analyzed. Therefore, we decided to rank the spectral components according to size, divide the ranks into a number of intervals, group the components within one interval and keep as original variables the energies of the spectral components in the groups. The number of original variables (i.e. the number of groups) was four or ten, depending on the number of analyzed items.

3. RESULTS

Figure 1 shows the first and second principal components describing the contour-free spectra of vowel [a] sustained by the pooled speakers (ten original variables.) One sees that the principal components assign the spectra characterized by different degrees of hoarseness to different chart zones. Figure 2 shows the evolution of the principal components of a (periodic) vowel signal whose cycle length has been increasingly perturbed and to which variable amounts of additive noise have been added. Increasing amounts of noise indeed move the coordinates that describe the spectra from the second to the fourth quadrant of the principal components plot.

A summary of the results obtained is as follows.

(a) For the contour-free spectra of vowels [a] [i] and [u] (pooled speakers, ten original variables), the correlations obtained between the original variables and the principal components were the same This means that the principal component charts were very similar too. When the number of original variables was four, and separate analysis were made both of the male and female speakers, and the healthy or dysphonic speakers, the correlation patterns obtained were also the same.

(b) The first principal component alone explained 97, 95, and 97 percent of the variability of the contour-free spectra of vowels [a], [i] and [u] respectively (pooled speakers, ten original variables.) Similar percentages

were obtained when the number of original variables was four and the male or female, and healthy or dysphonic speakers were analyzed separately.

(c) When the spectral contours were not removed, the first principal component explained 56 and 68 percent when the number of original variables were ten and four respectively (vowel [a], pooled speakers). The corresponding percentages that were explained by the second principal component were 27 and 24.

(d) The commonplace approach to spectral smoothing is cepstral analysis [10]. We carried out comparisons between spectral and wavelet analysis by means of synthetic vowels [a], [i] and [u] perturbed by different amounts of additive and modulation noise. We estimated the values of the first two formants via the prominent peaks of the spectral contour. The number of formant estimates was 150 (5 levels of modulation noise x 5 levels of additive noise x 3 vowel timbres x 2 formants.) The results showed that the average error of the formant frequencies was the same for the wavelet and cepstral transforms and that wavelet analysis and cepstral analyses were more precise in 74 and 76 cases respectively.

4. DISCUSSION AND CONCLUSION

(a) Examining the weights of the original variables in the principal components shows that hoarse voices are expected to be positioned in the fourth quadrant of the principal components plot (i.e. the chart obtained by means of the first two principal components.) Comparisons both with the degrees of hoarseness assigned by the panel and simulations show this to be the case (Figures 1 and 2).

(b) In the principal components plot adjacent degrees of hoarseness tend to overlap (Figure 1.) Examining pairs of voices shows that the overlap is a consequence of the approximate nature of the classification carried out by the panel. This may have been due to the fact that the judges had to classify contoured spectrograms sequentially. By contrast, the principal components analysis was performed on the contour-free spectra in toto. Informal listening tests of paired spectra showed that the first principal component had indeed ranked the contour-free spectra according to hoarseness. This indicates that hoarseness was the main source of inter-spectral variability.

(c) The first principal component explained more than 95 percent of the inter-speaker variability whatever the speaker group, vowel timbre or number of original variables. This demonstrates that the noise in the spectral data could be summarized via a single cue.

(d) Since the principal components were linear combinations of log-energies, their mathematical form was the logarithm of a ratio of the energies of spectral components grouped according to size. Principal components analysis therefore yielded cues that were formally similar to conventional ones involving ratios of the harmonic and inter-harmonic energies.

(e) When the contour was not removed prior to principal components analysis, the total variability could not be explained by a single principal component and the distinction between the hoarse and clear voices was fuzzy in the principal components plot. The reason was that the first two principal components explained less of the total variability, which was greater because it included the irrelative contour-related inter-spectral variability. The boosting of the inter-harmonics by the contour mixes inter-harmonics and harmonics, as well as inter-harmonics of different intrinsic sizes. Therefore, the energy of the (contoured) spectral components grouped according to their size does not vary monotonically with noise.

(f) Examining the numerical values of the first original variable - the one that was equal to the spectral energy of the biggest spectral components - showed that the numerical value changed by only two percent when the spectral components were grouped according to size in four instead of ten groups. This demonstrates that the harmonics did indeed determine the numerical value of the first original variable and that ranking was indeed able to segregate in practice the harmonics from the rest.

(g) We used wavelet analysis of the logarithmic vocoid spectrum to remove the spectral contour. We deemed the wavelet transform appropriate for separating the contour from the rest because, within vocoid spectra, the contour, and the noise and harmonics vary slowly and rapidly with frequency respectively. Principal components analysis attested to the fact that our choice was well founded, since it did not reveal any differences between vowel timbre [a], [i], and [u], thus indicating that the removal of the contour was successful.

REFERENCES

[1] De Krom G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. J. Speech and Hearing Research, 36, 254-266.

[2] Hiraoka N., Kitazoe Y. , Ueta H. , Tanaka S. and Tanabe M. (1984). Harmonic-intensity analysis of normal and hoarse voices. J. Acoust. Soc. Am., 76, 1648-1651.

[3] Kasuya H. and Ogawa S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. J. Acoust. Soc. Am., 80, 1329-1334.

[4] Michaelis D., Gramms T., Strube H. (1995), Glottal-to-noise excitation ratio – a new measure for describing pathologic voices. Acustica, 81, 700-706.

[5] Yumoto E. and Gould W. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. J. Acoust. Soc. Am., 71, 1544-1549.

[6] Qi Y. and Hillman R. (1997). Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. J. Acoust. Soc. Am., 537-543.

[7] Yanagihara N. (1967). Significance of harmonic changes and noise components in hoarseness. J. Speech and Hearing Research, 10, 531-541.

[8] Newland D. E. (1993). An Introduction to Random Vibrations, Spectral & Wavelet Analysis, Longman Scientific and Technical, Harlow, 295-370.

[9] Woods A., Fletcher P., Hughes A. (1986). Statistics in Language Studies. Cambridge University Press, Cambridge, U.K., 154-174.

[10] Verhelst W., Steenhout O. (1986), A New Model for the Short-Time Complex Cepstrum of Voiced Speech, IEEE Transactions on Acoust. , Speech and Sig. Proc., 34, 1, 43-51.