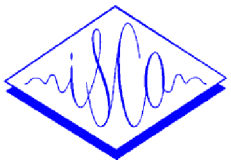


Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

WAVELET ANALYSIS OF SUSTAINED VOWEL SPECTRA IN VIEW OF THE CHARACTERIZATION OF HOARSENESS

Jean Schoentgen* & Fabrizio Bucella**
Laboratory of Experimental Phonetics
Université Libre de Bruxelles
Belgium
email: jschoent@ulb.ac.be, Fabrizio.Bucella@ulb.ac.be
*National Fund for Scientific Research, Belgium
** Fonds pour la Formation par la Recherche dans l'Industrie et l'Agriculture, Communauté Française de Belgique

ABSTRACT

This article concerns the spectral analysis of vowels sustained by hoarse and healthy speakers. Conventional cues of noise in spectral data are founded on the segregation of harmonics and inter-harmonics. What we propose here is an alternative that consists of flattening the spectral contour and performing a multivariate statistical analysis of the residual (i.e. flattened) spectral components. The residual spectrum is obtained by means of the wavelet transform of the power spectrum of the vowels. The results show that a principal components analysis of the flattened spectra enables hoarse voices to be separated from clear.

INTRODUCTION

Hoarseness is the overall description for a perceived deviation from the normal voice. Hoarseness is possibly the outcome of transients, unequal glottal pulses, aperiodic glottal waveforms, or increased disturbances of the glottal pulse amplitudes or lengths owing to jitter, shimmer and additive noise.

Noise in spectral data designates the patternless or unwanted components in the power spectra of vowels or other resonants. The usual descriptors of noise in spectra are founded on the segregation of harmonics and inter-harmonics, as well as on the calculation of a ratio involving the spectral energy of either the harmonics or inter-harmonics, and the total energy [1].

The rationale behind this approach is the following. Let us designate as the spectral contour a smooth line drawn through the spectral base-line, i.e. the many small patternless spectral components from which the harmonics emerge at low frequencies. Similarly, let us designate as the harmonic envelope a smooth line drawn through the tops of the harmonics. In the power spectrum of a periodic speech signal the spectral contour and the harmonic envelope would develop their dependent signal–shape profiles in parallel However when the signal is not periodic, the envelope and contour profiles are not equidistant. Instead, the harmonic envelope approaches the contour at higher frequencies. Indeed, the size of the harmonics decreases and that of the inter-harmonics increases with frequency. The gap between the spectral contour and the harmonic envelope is therefore a cue as to the noisiness of a spectrum, and calculating the ratio of the harmonics and inter-harmonics is akin to comparing the profiles of the envelope and the contour since the harmonics sample the former and the inter-harmonics the latter. The ratio of the sum of the squared magnitudes of the harmonics and inter-harmonics therefore constitutes a numerical cue that evolves monotonically with the noise in the spectrum.

Possible problems with this approach are as follows. Firstly, detecting the harmonics and segregating them from the rest of the spectrum is a task that cannot be undertaken statistically when the noise is not additive. Secondly, for a given level of modulation noise, the magnitudes of the harmonics decrease and their widths and fragmentation increase with increasing frequency until they blend into the other spectral components. This blending occurs gradually rather than abruptly. Consequently, the decision whether to assign a spectral component to the harmonics or to the noise is arbitrary up to a certain point. Thirdly, the spectral contour is related to the signal shape, which is an inter and intra-speaker variable even for the same vowel timbre. The discussion in the previous paragraph in fact suggests that the need to calculate a ratio arises only because the profile of a contour is not flat.

The rationale explained in the previous paragraphs suggests the replacement of the segregation of the harmonics and inter-harmonics by the flattening of the spectral contour. The harmonic envelope is in fact conditioned by both the spectral contour

and the noise, whatever its source. Flattening the contour therefore removes shape-related variability. As a consequence, the size of the residual spectral components is dictated only by the noise and whether a spectral component is positioned near a integer multiple of the average cycle length. For periodic signals, the flattened spectrum would therefore consist of harmonics of constant size and inter-harmonics that are zero or negligible. For aperiodic signals, the sizes of the (residual) harmonics decrease and the sizes of the (residual) inter-harmonics increase with frequency the more the further the signal is removed from strict periodicity. Consequently, the computation of the ratios can be omitted because the magnitudes of the residual spectral components depend on the noise only. A statistical analysis of the size of the residual spectral components will therefore reveal noisy spectra directly.

In practice, we ranked the spectral components according to their sizes, grouped those of similar size together and performed a principal components analysis of the energy of the grouped components. The number of groups was typically four or ten. The results show that principal components analysis is able to separate clear from hoarse voices and array increasingly perturbed synthetic signals.

METHODS

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 St Pierre University Hospital in Brussels. The pathologies, diagnosed by the ENT department's doctors, fell into the following categories: oedema of the vocal folds; nodules; hypotonia and asthenia; pseudo-cysts; granuloma; congestion of the vocal folds; polyps; chronic laryngitis; hyperkinesia and paralysis. The sampling frequency was 20 kHz.

Synthesis

Synthetic vowels [a], [i] and [u] were obtained by means of a formant synthesiser driven by a glottal pulse model. The lengths of the synthetic glottal pulses were perturbed in increments from 0% to 1%. The length disturbances were distributed normally. Uniformly distributed noise was, in addition, added to the glottal pulse during the open interval of the synthetic glottal pulse. The signal-to-noise ratio of the glottis signal was calculated to characterise the amount of added noise. The sampling frequency was 10 kHz.

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 a scheme proposed by Yanagihara [2].

Signal processing

The signal processing steps were as follows.

i) The normalisation of the signal energy and the Hamming windowing of a sustained vowel segment 16384 samples long. The sampling frequency was 10 or 20 kHz.
ii) The performance of the Fast Fourier Transform and the calculation of the power spectrum.
iii) The performance of a fast, discrete and invertible wavelet transform of the power spectrum.
iv) The assignment of the first 128 wavelet coefficients (constant included) to the spectral contour, and of the remaining 16256 to the residual (i.e. flattened) power spectrum.
v) The acquisition of the residual power spectrum by means of the inverse wavelet transform of the 16256 wavelet coefficients.
vi) The representation of the residual power spectrum in doubly logarithmic co-ordinates
vii) The fitting of a straight line to the baseline of the doubly logarithmic residual power spectrum
viii) The subtraction of the fitted line to obtain the flat power spectrum.

The removal of the contour (step v) in fact produced a residual power spectrum whose average was zero, and whose spectral baseline consequently dipped slightly below zero to offset the harmonics projecting above the baseline at low frequencies. The correction (step viii) of the positive tilt of the baseline was minor and could be omitted when the spectral resolution was high (because the harmonics were narrow).

Below we detail the wavelet analysis of the power spectrum, which is an alternative to cepstral smoothing. Wavelet analysis is in fact a multi-resolution analysis of a signal and basically consists of expanding the signal by means of oscillating waveforms (called wavelets) whose

average is zero and which differ significantly from zero on a finite interval only. The independent variables are the wavelet position and the length of the period during which the wavelet is different from zero (i.e. the wavelet scale). Because the number of oscillations is fixed, the wavelet in fact oscillates slowly on long, and rapidly on short, wavelet supports. As a consequence, the projection of a one-dimensional signal on to a one-dimensional wavelet is expected to be small when the rate of change of the signal at a given position is either faster or slower than the typical rate of change of the wavelet. On the other hand, the projection is expected to assume high values when the rates of change of wavelet and signal are similar. When the positions of the wavelets along the signal and their scales (fine or coarse) are fixed by the experimenter, the wavelet representation may be prolix or sparse. The outcome is an analysis that is non-invertible because the relation between the wavelet coefficients and the signal samples is not one-to-one [3].

A possible solution is to resort to a set of wavelets which form an orthogonal base. Daubechies wavelets, for instance, are orthogonal, zero beyond a finite interval, and almost smooth [4]. A fast, discrete, and invertible wavelet analysis thus consists firstly of the selection of a canonical wavelet via the number of the wavelet filter coefficients. The greater the number of filter coefficients, the smoother and broader the wavelet and the narrower its compass in the dual domain will be. We carried out experiments with wavelets defined via different numbers of filter coefficients. The results suggest that Daubechies wavelets with at least 12 filter coefficients are appropriate for the analysis of purely voiced speech or spectra. We selected a 20-coefficient Daubechies wavelet because it was well situated in the dual domain, with sidelobes of small magnitude. Secondly, we have to select a number of signal samples which is a power of two. This condition was automatically satisfied when the wavelet transform was applied to the log-power spectrum, which was 16384 samples long. The total number of wavelets was therefore equal to 16384, and the number of scales equal to 14 because $2^{14} = 16384$. The wavelets were automatically distributed among the 14 scales as follows. 2^{13} on the finest, 2^{12} on the second finest, and so on up to 2 on the coarsest. This means that on the finest scale, the wavelets were positioned two samples apart, on the second finest four samples, and so on. For any given scale, the positions of the wavelets were equidistant.Thirdly, a fast wavelet transform has to be performed which yields a number of coefficients equal to the number of samples. Signal constituents can be re-synthesised scale by scale and, together, reproduce the original exactly. Separating constituents consists of reproducing signal constituents by means of subsets of scales and processing or storing them separately. The residual spectrum was thus obtained by performing the inverse wavelet transform via the wavelets in the upper seven scales out of a total of fourteen. Indeed, the coarse wavelet scales were expected to represent the contour because the contour profile evolved slowly with frequency while the spectral noise and harmonic profiles did so rapidly.

We carried out experiments that showed empirically that this was the case, and that the segregation was optimal when the constant wavelet and the wavelets in the first seven scales were assigned to the contour, and the rest to the residual spectrum [5][6]. This result can be confirmed by reasoning. Let us assume that formants 200 Hz apart must be resolved within a spectral contour. The gap of 200 Hz corresponds to 164 spectral samples. According to Newland [7], the resolution expected within a wavelet scale is given by the frequency of repetition of the shifted wavelets. On scale number seven the wavelets were 128 spectral samples apart. This interval was the nearest to the one desired (i.e. 164), so scales number one to seven were adequate for representing the spectral contour at the desired level of detail.

Statistical analysis

We performed a principal components analysis of the squared magnitudes of the flattened spectral components. Principal components analysis transformed the original variables into new ones that were uncorrelated and accounted for decreasing proportions of the variance in the data. The new variables, i.e. the principal components, were defined as linear combinations of the original ones [8].

Ideally, the original variables would have been the spectral magnitudes. In practice, however, the number of original variables must not exceed a small fraction of the number of items (i.e. spectra) analysed. Therefore, the spectral components were ranked according to size, the ranks divided into four or ten intervals, and the spectral energies of the grouped components were kept as the original variables.

RESULTS

a) Figure 1 shows the first two principal components of the spectra of synthetic vowels to which uniformly distributed noise was added during the open phase of the synthetic glottal pulse. Spectrum 1 has a glottal pulse-to-noise ratio of infinity (no noise) and signal 25 a ratio of 29.04 dB. One sees that increasingly noisy spectra are arrayed sequentially from right to left.

b) Figure 2 shows the first two principal components of the spectra of sustained synthetic vowels whose glottal cycle lengths were randomly perturbed in the range from 0 to 5 percent. The perturbations were distributed normally. The length perturbation of signals number 1 and 25 were 0. and 4.60 percent respectively. One sees that increasingly noisy spectra were arrayed sequentially from right to left. The number of original variables was four, i.e. the ranked spectral components were shared out among four groups.

c) Figure 3 shows the result of the principal components analysis of the spectra of the [a]-vowels sustained by 81 speakers. The number of original variables was ten. One sees that the first two principal components assigned spectra characterised by different degrees of hoarseness to different chart zones.

d) We also performed principal components analysis of the flattened spectra of vowels [i] and [u]. The correlation obtained between the original variables (ten) and the two principal components was the same for vowels [a], [i] and [u]. This means that the principal component charts were very similar.

e) The observation under the (d) above attests to the fact that the spectral contour analysis was well founded because principal components analysis of the residual spectra did not reveal any differences between vowel timbres [a], [i] and [u], thus indicating that the flattening of the contour was successful.

f) Informal listening tests confirmed that the overlap between the degrees of hoarseness observed in Figure 3 was the consequence of the inability of the panel to separate the different degrees of hoarseness crisply. Indeed, a two-by-two auditory comparison showed that the principal components analysis had arrayed the voices more reliably than the human judges.

g) The first principal component explained more than 95 percent of the inter-speaker variability whatever the speakers' groups (healthy or dysphonic), and genders, the vowel timbre or the number of original variables (four or ten). This confirmed that the conventional calculation of a single numerical cue is able to summarise the amount of noise in spectra.

h) The conventional approach to spectral smoothing is cepstral analysis [9]. We compared the cepstral and wavelet methods of obtaining the spectral contour. We therefore estimated the values of the first two formants via the local peaks of the spectral contours of 150 synthetic vowels that had been disturbed by variable amounts of additive and modulation noise. The results showed firstly that the average error of the formant frequencies was the same for the cepstral and wavelet transforms and secondly that wavelet and cepstral analyses were more precise in 74 and 76 cases respectively.

REFERENCES

[1] Kasuya H. and Ogawa S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. Journal of the Acoustical Society of America, 80, 1329-1334.

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

[3] Rao R., Bopardikar A. (1998). Wavelet Transforms, Addison-Wesley Longman Inc., Reading, Mass.

[4] Daubechies I. (1988), Orthonomal bases of compactly supported wavelets, Comm. Pure and Applied Mathematics, 41, 909-996

[5] Bensaid M., Schoentgen J., Ciocea S. (1997). Estimation of the formant frequencies by means of a wavelet transform of the speech spectrum, Proceedings 8th annual ProRISC & IEEE-Benelux Workshop on Circuits, Systems and Signal Processing, (pp. 49-54). Utrecht: STW-Technology Foundation, The Netherlands

[6] Micallef P., Chilton E. (1995). Spectral envelope of speech using wavelets, Proceedings EUROSPEECH 95, Madrid, Spain, 251-254

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

[8] Woods A., Fletcher P., Hughes A. (1986). Statistics in Language Studies, Cambridge: Cambridge University Press

[9] Deller J. R., Proakis J. G., Hansen J. H. L. (1993). Discrete-time processing of speech signals. New York: Macmillan Publishing Company

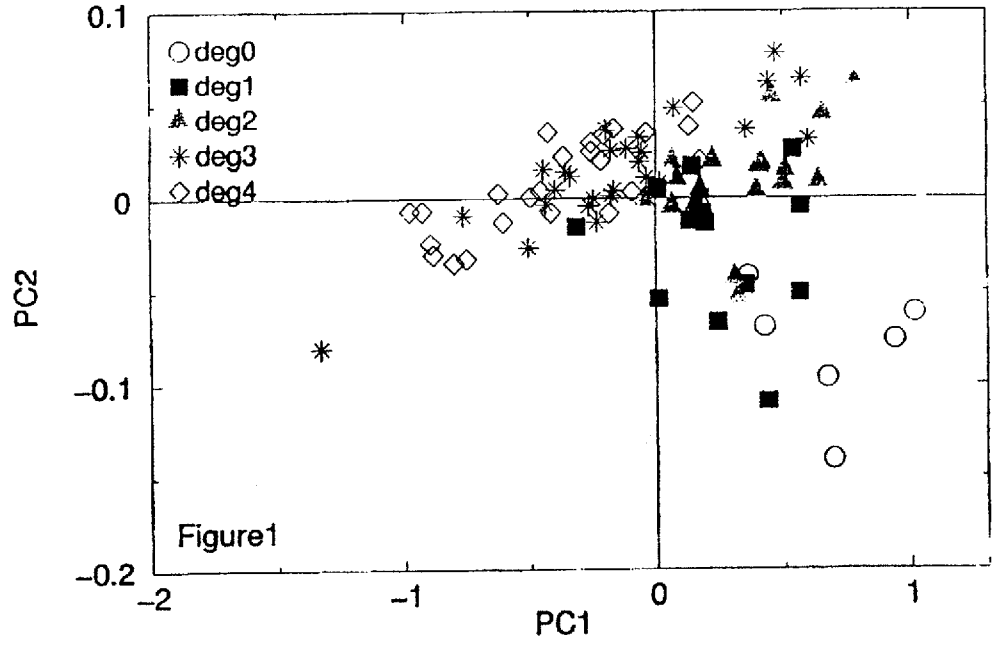


Figure 1: additive noise

Figure 2: modulation noise

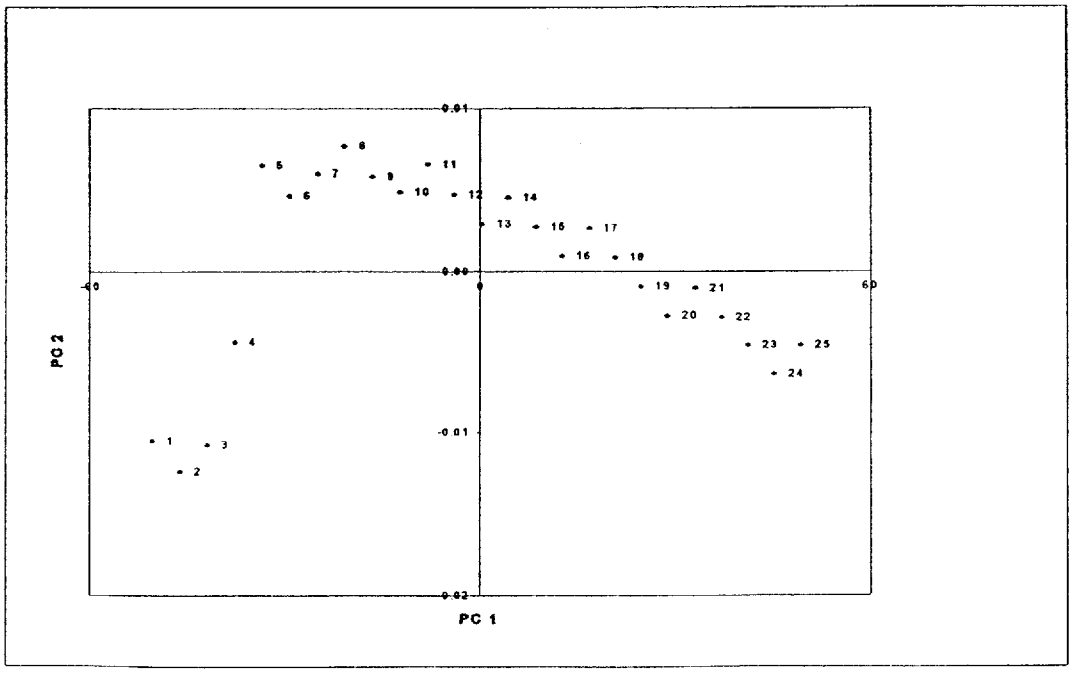
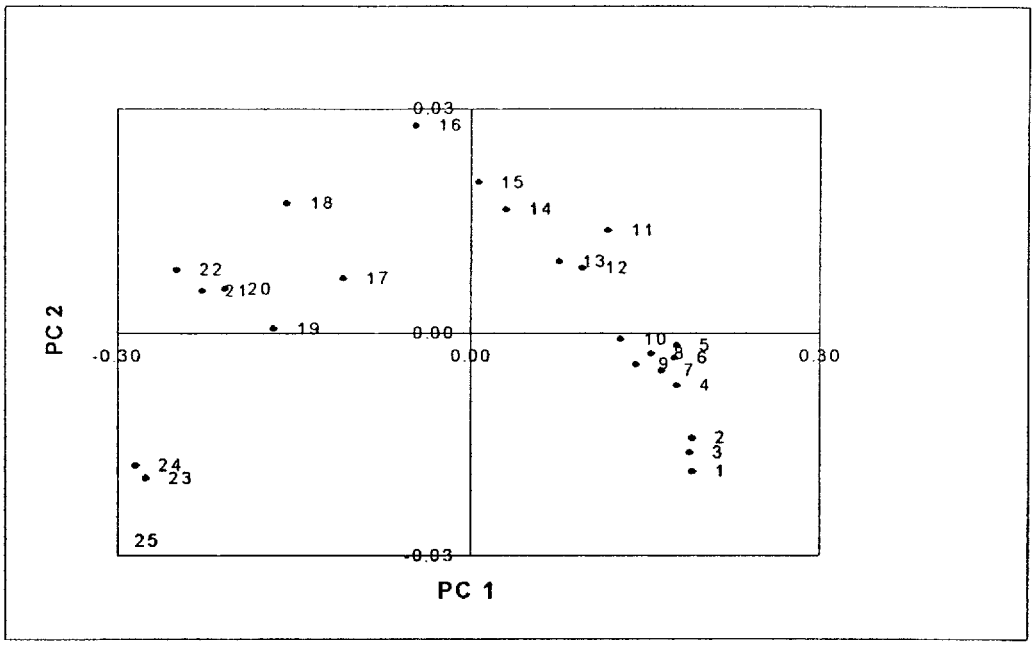


Figure 3


0.03
16
15
18
14
11
13
12
22
17
21
20
19
0.00
10
9
8
7
6
5
4
-0.30
0.00
0.30
2
3
1
24
23
25
-0.03
PC 2
PC 1