THE WAVELET TRANSFORM AND PHONETIC ANALYSIS

P.Basile - F.Cutugno - P.Maturi

Centro Interdipartimentale di Ricerca
per l'Analisi e la Sintesi dei Segnali
Università di Napoli "Federico II"

ABSTRACT - The advantages of the application of a wavelet-transform analysis system to spoken materials are here discussed, with reference to some Italian examples.

INTRODUCTION

Acoustic phonetics has so far been practised through spectro- and sonagraphical measurements based on the Fourier analysis, or Short Time Fourier Transform (STFT). An STFT analysis consists in filtering the signals through a bank of band-pass filters with constant bandwidth. The result is the separation of the harmonics when the filters' bandwidth is adequately lower than the fundamental frequency (F0), or an estimate of the formant structure when the bandwidth is higher than F0 (Portnoff 1980, 1981).

Audiological research has shown on the contrary that the human ear does not analyse acoustic signals in the same way as our sonagraphs do. The ear behaves in fact as a bank of filters of different bandwidths, so that the higher the central frequency of a filter, the wider its band. The result is a very accurate separation of lower frequencies, which allows us to detect the fundamental frequency and the first harmonics, and at the same time a rougher analysis at higher frequencies, where we only need to perceive formants (Moore, 1989).

An attempt to imitate this kind of analysis can be made by using the wavelet transform (WT). The WT consists in an analysis of the signal which is performed by a bank of filters having a constant *bandwidth/central frequency* ratio. This system approaches satisfactorily well the auditory analysis, even though the ear does not show such a perfectly constant ratio at all frequencies as the artificial systems do.

The application of a WT-based analysis to phonetic signals can thus yield a representation which allows us to evaluate the perceptually relevant features of the signal. Moreover, a WT analysis results in a single representation of the signal where both harmonic and formant structures can be simultaneously observed and where discrimination can be very accurate both in frequency and in time.

Notwithstanding its 'realism' and its practical advantages, the WT hasn't found a wide application in phonetics yet, nor are there any WT-based (or hearing-based) analysis systems on the market.

In the next paragraphs we will describe a particular WT-system we have implemented in our labs and will discuss the results of some test analyses performed on Italian sequences.

THE WT SYSTEM

Our analysis system is based on an integral representation of the wavelet transform. The basic idea of the WT is that a non-stationary signal can be expressed as a superposition of elementary components whose localization depends on a scale parameter (Portnoff, 1981; Kronland-Martinet & Grossman, 1991).

Formally, the WT results in a time-scale, rather than in a time-frequency (Cohen, 1989), representation and is defined by

$$T(\tau,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} g^*\left(\frac{t-\tau}{a}\right) x(t)\, dt \qquad (1)$$

where g*(t) stands for the complex conjugate of g(t), called the analyzing wavelet, whose constraints are described in (Kronland-Martinet & Grossman, 1991).

The elementary components in (1), as one can see, all derive from a mother function g(t) by translation and dilatation:

$$g_{\tau,a}(t) = g\left(\frac{t-\tau}{a}\right)$$

It can be shown that $g^*_{\tau,a}(t)$ defines a band-pass filter whose bandwidth depends on the dilatation parameter *a* in such a way that Q (i.e. the bandwidth/central frequency ratio) is a constant of the transformation.

The analysing wavelet we have used for the present work has the following form:

$$g(t) = h(t)e^{j\omega_0 t}$$

where h(t) is the so called Hanning window. In this form the relationship between the dilatation parameter *a* and the central angular frequency ω of $g^*_{\tau,a}(t)$ is simple.

In this work the definition of the temporal support of g(t) is such that we obtain a $Q \cong 0.33$, similar to the behaviour of the ear in a wide spectral range (i.e. above 400 Hz).

In dependence of the constant-Q property, the Wavelet transform shows a better frequency resolution in the lower part of the spectrum, where it is capable to separate the first harmonics (generally at least the first two, as you can see in the pictures shown), and a better time resolution in the upper part, so that the formants are clearly shown with no superimposition of the harmonic structure (Basile et al., 1992).

The WT can thus be also interpreted as a bank of filters, whose bandwidth grows together with the frequency.

As a consequence of its properties, described above, we think that the WT, if applied to phonetic analysis, can offer at least three very important advantages:

1) it gives information both on the harmonic and on the formant structure in the same sonagram, whatever the pitch range of the signal; the user does not have to make any a priori hypotheses about the pitch, while, when using a STFT system, he is obliged to choose previously a particular window length and to study narrow-band and wide-band sonagrams separately;

2) it yields a representation of the signal where the auditorily relevant characteristics are shown;

3) a good frequential localization (in the lower part of the spectrum) and a good temporal localization (in the upper part of the spectrum) are both possible in the same sonagram.

MATERIALS AND METHODS

We have used the above described system to analyse some sequences uttered by a standard Italian male speaker. In order to have a sample of all Italian phones, we have built up two very simple lists of short sequences (most of the words included are nonsense, though some of them, accidentally, are real words). The list is the following:

- ['apa, 'aba, 'ata, 'ada, 'aka, 'aga, 'afa, 'ava, 'asa, 'aza, 'aʃʃa, 'attsa, 'addza, 'atʃa, 'adʒa, 'ama, 'ana, 'aɲɲa, 'ala, 'aʎʎa, 'ara, 'aja, 'awa], that is all 23 Italian consonants in an ['aCa] context;

- ['tita, 'teta, 'tɛta, 'tata, 'tɔta, 'tota, 'tuta], that is all 7 Italian vowels in a ['tVta] context.

All signals were sampled at 20480 Hz with a 16-bit ADC and then processed off-line by a host computer equipped with a DSP coprocessor board.

In all the produced graphs the amplitude is shown by means of a grey-level scale, just like in traditional sonagrams. The x-axis corresponds to time; the y-axis displays frequency along a log scale.

SOME OBSERVATIONS

We will summarize here some of the most interesting results of the analyses performed with our system on the materials listed in the previous paragraph.

- Voiced/voiceless: what we usually call "voice bar" appears here in very good detail as a group of two or three harmonics (see Fig.1a-b);

- explosion: the burst can be observed very clearly, and so can its temporal distance from the first vowel period (VOT) (see Fig.1a);

- noise: fricative noise is detected both at low and at high frequency (see Fig.2a-b); a voiced fricative also shows harmonics; [v], on the contrary, doesn't show any noise at all, and has many harmonics and even a formant structure, while its voiceless counterpart, [f], has a very intense low-frequency noise (see Fig.3a-b);

- vowels: due to the log scale, the formant structure of vowels generally appears in the upper part of the graphs (Fig.4a-b); although the analysis scale is not linear, one could think of a linear-scale representation, which would allow a better visualization of the formant structure;

- pitch: although the sequences have been spoken in isolation with no sentence intonation (apart from the list-reading effect), the speaker's pitch shows spontaneous variations, which are very clearly observable in all graphs.

REFERENCES

Basile P., Cutugno F., Maturi P., Piccialli A. (1992) *The time-scale transform method as an instrument for phonetic analysis*, in: M.Cooke & S.Beet (eds.), *Visual Representations of Speech Signals*, (John Wiley & Sons: Chichester), in press;

Cohen L. (1989) *Time-Frequency Distribution - A review*, Proceedings of the IEEE, Vol. 77 (7), pp. 941-981;

Flandrin P. (1987) *Some aspects of non-stationary signal processing with emphasis on time-frequency and time-scale methods*, in: Comb J.M., Grossmann A., Tchamitchian Ph. (editors), *Wavelets: time-frequency methods and phase space,* (Springer-Verlag: Berlin), pp. 68-97;

Kronland-Martinet R., Grossman A. (1991) *Application of Time-Frequency and Time-scale Methods (Wavelet transforms) to the Analysis, Synthesis, and transformation of natural sound*, in De Poli G., Piccialli A., Roads C. (eds.), *Representation of Musical Signals* (MIT Press: Cambridge, MA) pp.43-85.

Moore B.C.J. (1989) *Introduction to the psychology of hearing* (Academic Press: London);

Portnoff M.R. (1980) *Time-frequency representation of digital signals and systems based on short-time Fourier analysis*, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 28(1), pp. 55-69;

Portnoff M.R. (1981) *Short-time Fourier Analysis of sampled speech*, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 29(3), pp.364-373.

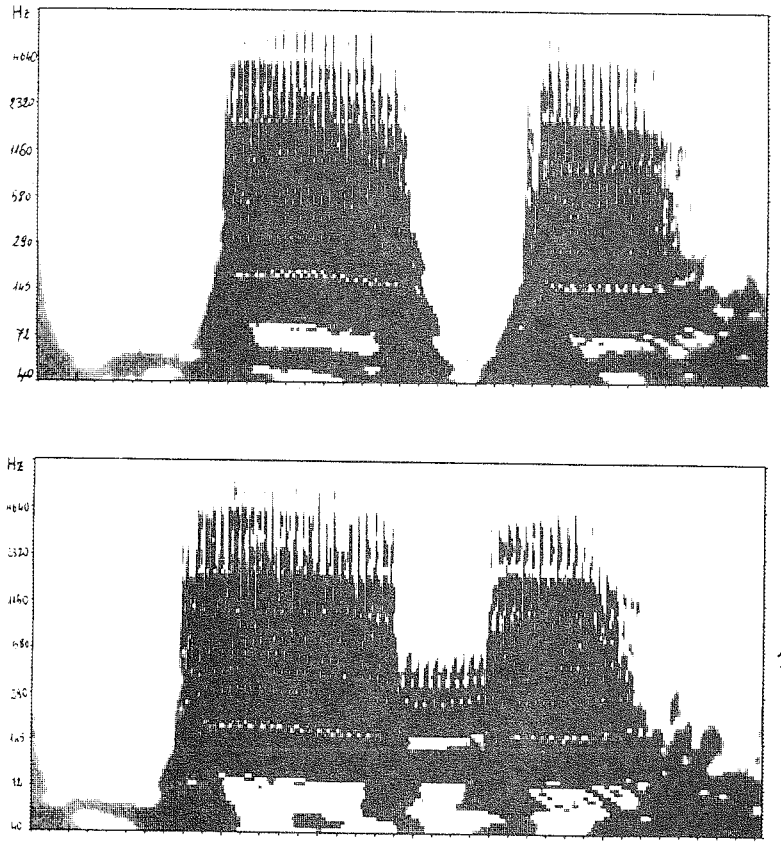
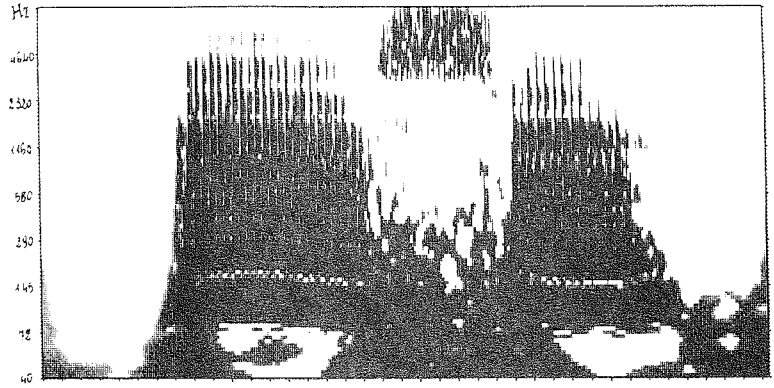
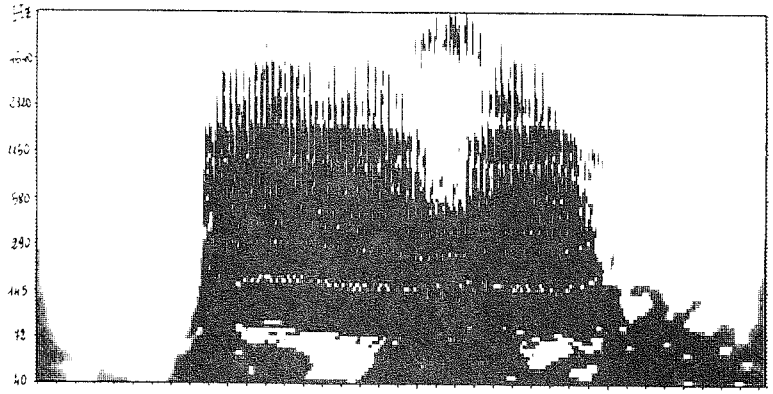


Figure1. WT sonagrams of the Italian sequences ['ata] (a) and ['ada] (b).

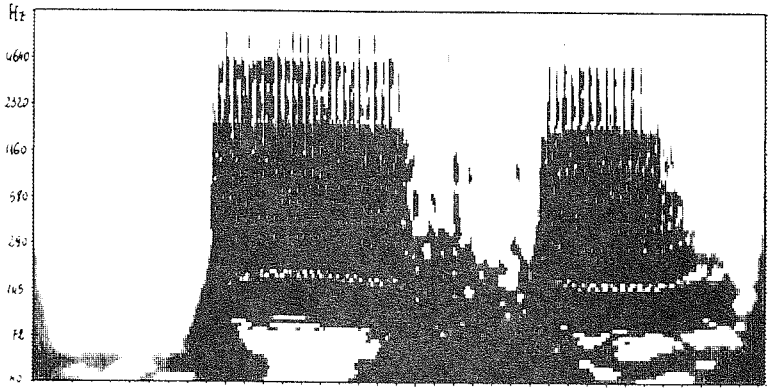


2a



2b

Figure 2. WT sonagrams of the Italian sequences ['asa] (a) and ['aza] (b).



3a

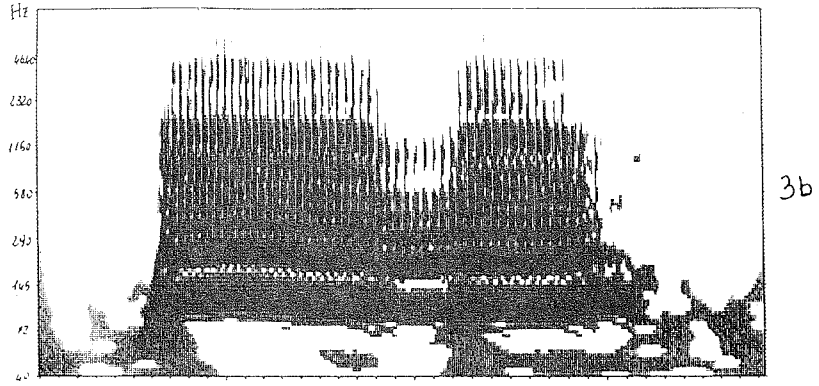


Figure 3. WT sonagrams of the Italian sequences ['afa] (a) and ['ava] (b).

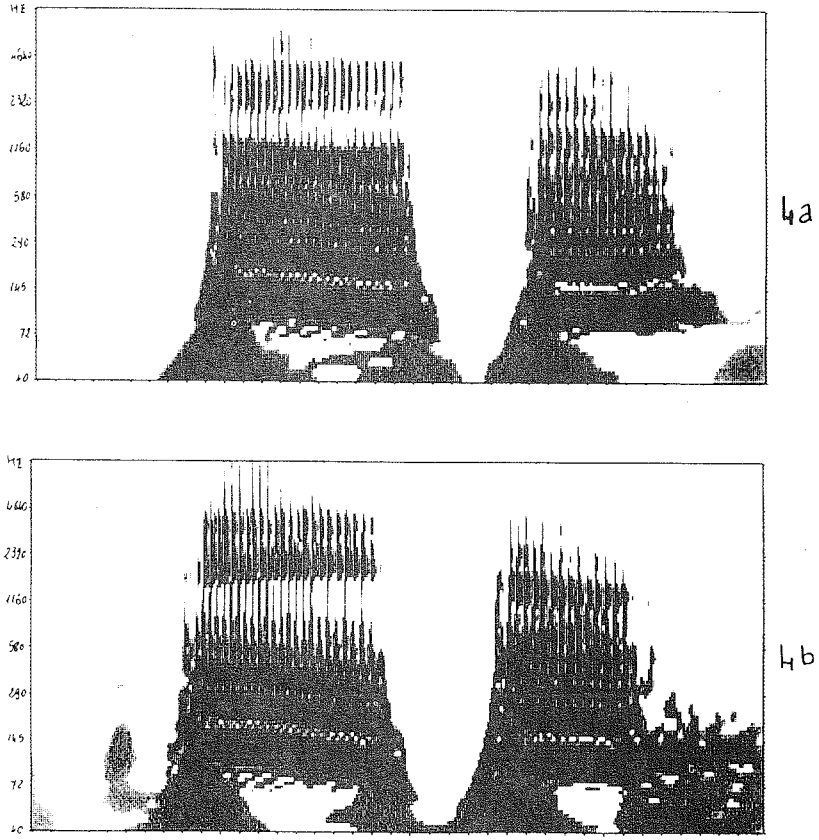


Figure 4. WT sonagrams of the Italian sequences ['tota] (a) and ['teta] (b).