Wavelets and Granular Analysis of Speech

J.S. Liénard and C. d'Alessandro
LIMSI-CNRS, BP 30, F-91406 Orsay Cedex, France

I - Very short time analysis of speech

The speech signal comes from the convolution of a source signal - due to the vibration of the vocal cords or to the airflow through a narrowing of the vocal tract - with the impulse response of the vocal tract. Both constituents rapidly change over time, and one usually considers that the phonetic information in the signal is mainly related to the evolution of the two or three first resonances of the vocal tract, called "formants" F1, F2, F3. The vocal cords vibrating frequency, Fo, is closely related to a perceptive quality of sounds called "pitch".

In order to extract the phonetic information the signal is considered to be stationary over a time interval long enough to include several pitch periods, and short enough to capture the evolution of the spectral envelope. The usual tradeoff yields some 25 ms for the width of the analysis window, and 10 ms for the interval between successive windows. So, despite the fact that everybody agrees about the relevance of classical spectrographic analysis - which uses bandpass filters 300 Hz wide, with impulse responses as short as a few ms - the information extracted from the signal for transmission or recognition is altered from the start in the time dimension. Some rapid transitions in consonants are smoothed or even erased, the sound structure disappears, the possible perceptual interaction between segmental and prosodic information is deliberately discarded.

The "granular" analysis we present hereafter aims at decomposing the signal into a set of discrete elements associated with energy concentrations in the time-frequency coordinates, with some emphasis on the time resolution (in the 1 to 2 ms range). In the voiced segments (vowels, some consonants) those grains correspond to the resonance maxima of each proper mode of the vocal tract, at each pitch period. In the noise segments (consonants such as "s" or "ch") the grains are randomly distributed in some region of the time-frequency plane. Finally the bursts found at the onset of some sounds like "p" or "t" give birth to one or several grains precisely located at the same instant, following a silence. In our view of speech analysis the notions of voicing, pitch, formants, for which no method gives a completely satisfactory answer, cannot be directly extracted from the signal, but should result from a structural study of the grain distribution. For instance the signal will be declared as "voiced" when, locally, comparable grains appear at regular intervals.

This analysis is based on hypotheses about the temporal coding of the acoustical wave by the human auditory apparatus. Consequently it is tempting to implement an auditory model in order to check them. However, it is difficult to validate and interpret the results of such models, because they implicitly take into account some further processing by the brain, which cannot be modeled or understood as yet. Thus we choose to implement the granular analysis in a way such that objective or subjective verification is permitted through a reconstruction process (ref 1).

The analysis process is composed of two steps. The first one aims at decomposing the signal into a series of narrowband signals, covering the frequency band of speech, i.e. from 70 to 5000 Hz. The second step consists in modeling each of them into a series of successive elementary waveforms. At the present time the first step only can be related to the theory of wavelet analysis.

2 - Decomposing the speech wave into a set of narrowband signals

In order to reconstruct the signal by a mere addition of its narrowband components it is necessary that all of the filters respond with the same phase, have the same slopes, and have their gains properly adjusted with respect to the distribution of the center frequencies along the frequency scale. We implemented a recursive filter structure, used twice with time-reversal in order to cancel any delay or phase distortion. The result is a zero-phase filter, the order of which is twice the initial filter order. For the basic unidirectional filter we choose a simple resonator (second order), so that the resulting filter is of order four (slopes at infinity tend toward -12 dB/octave, slopes around the cutoff frequencies depend on the quality factor, Q).

The distribution of the center frequencies along the frequency scale is one of the filterbank parameters. We used several tunings, ranging from linear to logarithmic, with an intermediary choice (Bark scale) close to what is known of ear physiology (tendency toward the linear scale in the low frequencies, toward the logarithmic scale in the high frequencies). The gains are automatically adjusted with respect to the number and distribution of the filters. Usually the number of filters is between 12 and 32, the bandwidths range from 100 to 600 Hz, the quality factor remains within the 1 to 10 range.

The filtering process is illustrated on Fig 1, which shows the decomposition of a series of impulses into 16 linearly distributed channels, as well as their additive reconstruction. Except for some noise due to the poor bandpass limitation of the signal, it is clear that the reconstruction is satisfactory.

Fig 2 shows the analysis of a speech signal, with some differences in the filterbank parameters, and a different representation of the output signals : only the positive parts of each signal are represented, after logarithmic compression of the amplitude. This repre-

sentation exemplifies the synchronization phenomena occurring among adjacent channels when several filters capture the same signal component. Here again, reconstructing the signal through summation of the outputs yields a signal very close to the original. When listening to both, no difference can be heard.

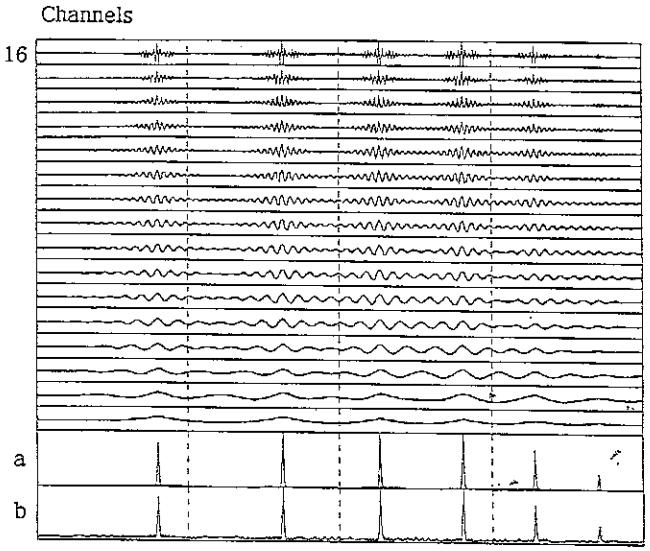

Fig 1 - Decomposition of a series of pulses (a) with a zero-phase filterbank, and signal reconstructed (b) by summation of the 16 output signals. Time scale 10 ms between vertical dotted lines (valid for all the figures in the present paper).

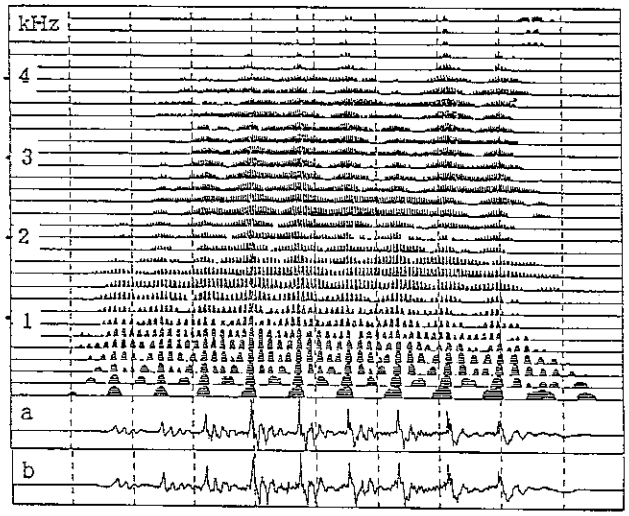

Fig 2 - Vowel "a" analysed with a 32-channel linear filterbank; a) original, b) reconstructed signal.

Basically our analysis process consists of the convolution of the signal with a symmetrical "wavelet" made of the bidirectional impulse response of the filter. Comparison with the Morlet and Grossmann wavelet theory yields some remarks, which can be classified as analogies and differences. The first similarity is to be found in the need to a better mastery the tradeoff between time and frequency resolutions. In both cases a better time resolution is expected in the high frequency part of the spectrum, while the frequency precision is expected in the low frequency part. Another important similarity lies in the additive reconstruction possibility.

As for the differences, the formal expressions given by the wavelet theory are obviously an advantage, thanks to the insights and guarantees they provide. On the other hand, the logarithmic frequency scale imposed by the wavelet shape conservation in its compressions and dilations may not be perfectly adapted to our psycho-physiological needs, and is not mandatory in our approach. In most of our experiments the equivalent "wavelets" composed of the bidirectional impulse responses of the filters were closer to the Gabor type than to the Morlet-Grossmann type. Finally the last noticeable difference lies in the implementation of a recursive IIR filter, which allows the computations to be achieved very quickly, several orders of magnitude faster than the wavelet analysis. The filter used has a low Q factor, and does not cause any stability problem.

3 - Modeling the output signals into discrete elements

If speech processing could be reduced to decomposing the signal into n narrowband signals, we would have gained nothing; the information rate of the signal would simply be multiplied by n. We are actually looking for a decomposition into a set of discrete elements, or grains, which we call Elementary Waveform Models, or wfms (fig 3). This decomposition has been validated for singing voice synthesis (ref 2), but our problem is the opposite, i.e. how to go from the signal to the list of grains ?

A first way to deal with this probleme is to decompose each narrowband signal into a string of wfms. For this we spot the extrema of the envelope, associate to each maximum a prototypical waveform whose parameters have been adjusted so that the sum of two successive wfms makes up a good approximation of the signal for the zone being

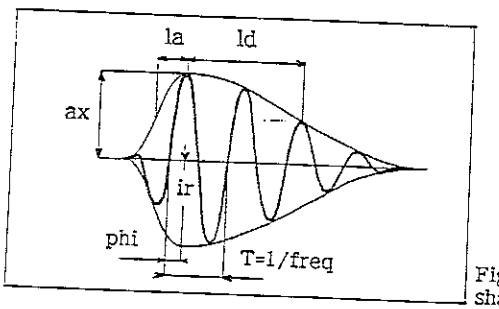

Fig 3 - Waveform model, with attack and decay shaped by raised sinusoids.

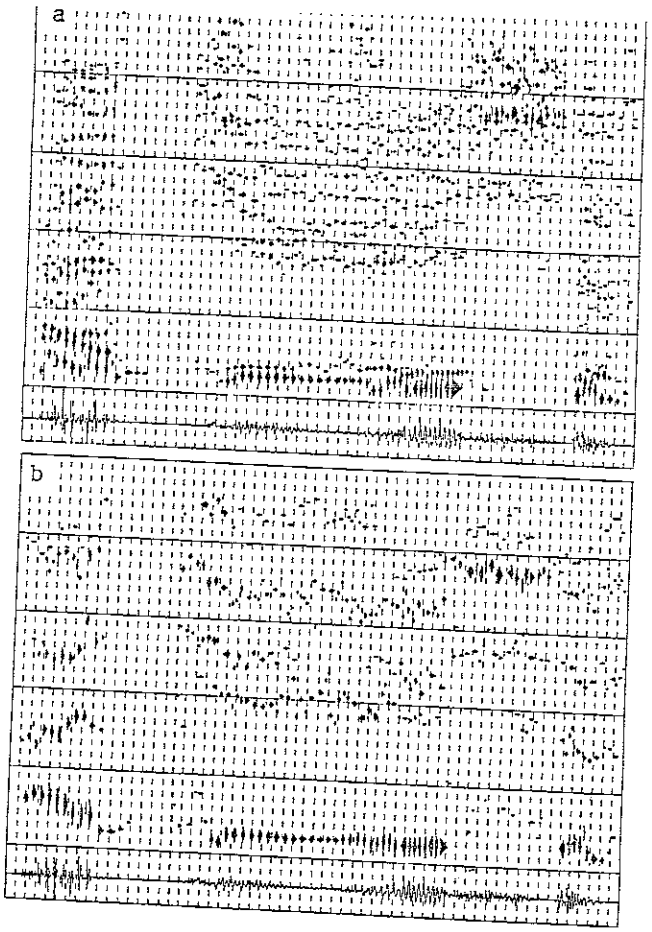

Fig 4 - Speech segment /atyvysə/ displayed as a set of wfms in the time-frequency plane. (a) : channel by channel modeling. (b) : same, after iterative grouping of adjacent channels.

modeled. This process produces a set of wfms for each channel (fig 4a). Reconstruction of the signal by regenerating the wfms and summation of all of the channels furnishes a signal perceptually very close to the original.

Even though this represents a considerable economy compared to the set of narrowband signals produced by the filterbank, the channel-by-channel modeling process still is highly redundant. As the filters we use are not extremely selective, one given constituent of the signal has been analysed by several filters, resulting in several wfms in adjacent channels. It is only when the sum of all of the wfms in those adjacent filters is calculated that that the original constituent - grain - can be regenerated. We have therefore created a local grouping procedure for the narrowband signals surrounding any local maximum of the envelope in the time-frequency space. This procedure greatly reduces the total

number of wfms obtained (fig 4b), and should ensure their invariance against different filterbank configurations, but some spotting problems appear in the low frequency range where the pitch period and the center frequency of the analysis channel are close to each other.

Since the grouping procedure described above is costly and not perfect, a third approach has been elaborated which makes use of the specific structure of the speech signal (ref 3). It consists of filtering a short segment of the signal (some 50 ms, in order to avoid any boudary effect) in the regions of spectral prominence, as evaluated by a classical LPC analysis, rather than in fixed, permanently defined frequency bands. The new regions are frequently reestimated (every 6 ms in our experiments). Each filtered signal is then segmented and modeled by the same procedure, giving the desired grains without the need of a grouping procedure. This approach gives satisfactory results if it is adapted to the lower part of the spectrum (modeling the first harmonics) (ref 4).

4 - Conclusion

Our short term analysis, as well as wavelet analysis, has the desire to dominate the compromise between time and frequency resolutions. Both processes may be seen as filtering, or as the convolution of the signal with a particular, symmetric, Gaussian shaped kernel. But beyond this common desire, we are trying to model the signal into a set of discrete elements, or grains, which are supposed to be perceptually pertinent. This aim contributes to defining an inverse problem for which, at present, wavelet theory has no answer.

5 - References

1 - J.S.Liénard : "Speech Analysis and Reconstruction Using Short-Time, Elementary Waveforms", IEEE-ICASSP, Dallas, 1987.

2 - X.Rodet : "Time-Domain Formant-Wave-Function Synthesis", Computer Music Journal, vol 8, 3, 1985.

3 - C. d'Alessandro and J.S.Liénard : "Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation", IEEE-ICASSP, New York, 1988.

4 - C. d'Alessandro : "Analyse-Synthèse de la bande de base par formes d'ondes élémentaires", 17e Journées d'Etude sur la Parole de la SFA, Nancy, 1988.