S8.13

# DECOMPOSITION OF THE SPEECH SIGNAL INTO SHORT-TIME WAVEFORMS USING SPECTRAL SEGMENTATION

Christophe d'Alessandro and Jean-Sylvain Liénard

LIMSI-CNRS, Orsay, France

ABSTRACT

Speech representation by a set of short-time, elementary waveforms appears as a new approach to speech processing. In the present study, the signal is pre-analysed frame by frame; the spectral envelope obtained for each frame is segmented into regions comprising a single peak. The signal is then filtered in each region, and the elementary waveforms are spotted in the time domain. The problem of grouping the waveforms in adjacent channels is thus circumvented. The resulting representation is satisfactory, as well as the signal reconstruction, except for some modelling problems remaining in the lowest part of the spectrum.

## I - Introduction

This paper continues our work (ref 1) on a representation of the speech signal by a set of discrete elements which respect its acoustical and perceptive structures.

First, the temporal resolution of the analysis is given more importance than in traditional analyses. This option is shared by a recently developped analysis method (wavelet analysis, ref 2), which, after GABOR's work, decomposes the signal into well-localized time-frequency energy concentrations.

Second, we want to define elements within the signal ("grains", or elementary waveforms wfs) that contain all of the perceptual information, without, at this level, defining voicing or pitch explicitely. The concept of elementary waveform is close to X.RODET's FOF (Formant Wave Function), which has been used successfully for high-quality synthesis of singing voices (ref 3).

The "granular spectrogram" that results from this analysis will later be used to look for the classic elements of perception acoustics : voicing, pitch, formants, bursts etc. In order to make sure of the analysis relevance, we use to validate the decomposition by resynthesis. At the present stage of our study, we do not try to turn it into a coding method.

In the process presented in ref 1, the short-term spectral envelope of the speech signal was obtained using a zero-phase filterbank. The grains were defined through channel-by-channel modelling. After resynthesis, the quality obtained was excellent, but the representation was still redundant. Local grouping of adjacent channels yielded the desired representation; however some quality problems were encountered in the lower part of the spectrum during modelling and resynthesis.

We present here a somewhat different method based on spectral segmentation before temporal modelling. In this manner, we try to profit from a specificity of the speech signal, which has a spectrum composed of peaks : the interval between two valleys corresponds to the number of adjacent channels that were to be grouped in the former processing.

In section II we explain the principle of the analysis-synthesis model. In section III the system developed according to this principle is described. Some results are presented in section IV.

## II - Overview and Production Model

In the traditional approaches (FFT, LPC, Cepstrum etc), one uses successive analysis windows, regularly spaced, of fixed length (often long enough to include several pitch periods). This way to perform the analysis tends to separate in the first place two different aspects of the signal: pitch (and the parameters related to the excitation), and spectral envelope. The excitation signal is then modelled, either in a rather crude way (the binary feature voiced/unvoiced is used in many systems), or more finely (for example, a sequence of pulses in the multipulse methods), but with a serious drawback, as far as interpretation capability is sought: there are no relations between the secondary pulses and the acoustic or perceptive structures.



CH2561-9/88/0000-0351 $1.00 © 1988 IEEE

As we wish to work out a method which provides a description of the signal preserving or emphasizing those structures, we will try to model the source and the spectral envelope in both spectral and time domains. The whole processing is summarized in fig 1.

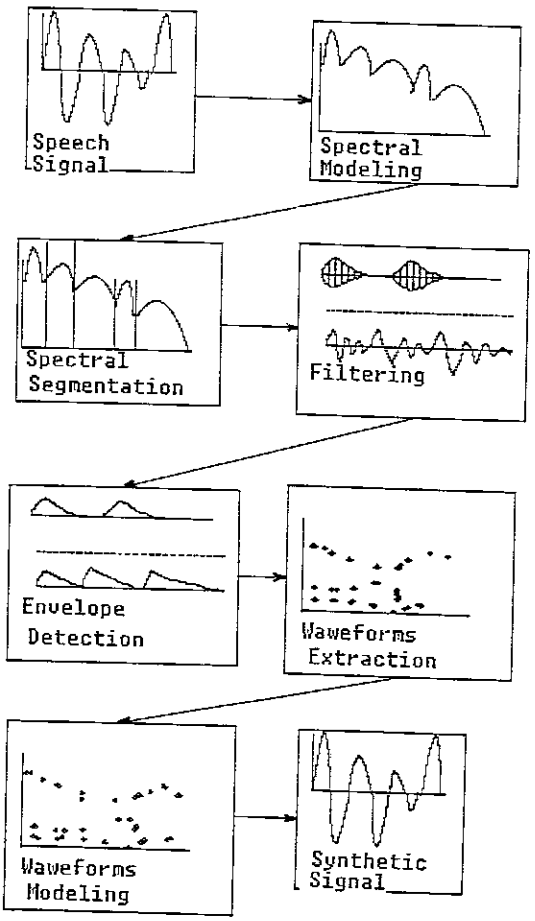


Fig 1 - The Analysis and Reconstruction Process.

The signal is first pre-analyzed, frame by frame, using a classical LPC algorithm; it is therefore modelled through an all-pole model. For each frame, the spectral envelope is segmented into regions, each containing one single envelope maximum. The signal is then filtered in each region, the elementary waveforms wfs are spotted between two successive minima of the time envelope, their parameters are evaluated (wf modelling), and reconstruction is achieved by summing the appropriate set of waveform models wfms.

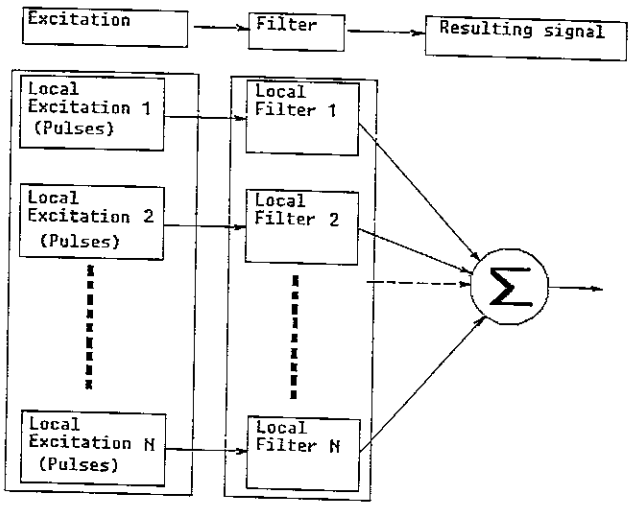


Fig 2 - The production model

An elementary waveform wf can be considered as the response of a spectrally local filter to a spectrally local excitation (fig 2). Under some conditions explained below, this local source can be modelled by a set of pulses, and the wfs can be identified with the impulse responses of the local filters - in the formant regions. Each "grain", or wf, represents a spectro-temporal event. The process can be formulated as follows.

$$s(t) = \sum_{i=0}^{N} \sum_{j=0}^{N_j} \alpha_{ij} \, ( \delta_{ij} * p_i * f)(t)$$

s(t): speech signal
N : number of formants
$N_j$ : number of pulses
$\alpha_{ij}$: amplitude
$p_i$ : impulse response of the spectral segmentation filter
$\delta_{ij}$ : unit impulse at instant $\tau_{ij}$
f : impulse response of filter F given by the production model
* : convolution operator
i : index of formant bands: $1 \leq i \leq N$
j : index of the pulses in band i: $0 \leq j < N_i$

It should be mentioned that the model presented includes the classical source model, as well as the multipulse model. If a pulse appears exactly at the same time in all analysis bands, it provides a single pulse exciting the global filter F. We thus endorse the hypothesis - successfully used in multipulse analysis - that all types of excitation can be viewed as sequences of pulses. We just consider here a sequence of pulses to be localized in each formant region, in order to give a better account of its acoustical and perceptual contribution.

## III - Experiments

Our method of analysis-synthesis allows for detection, modelling, and resynthesis in the time domain, of a set of elementary waveforms which will be defined below. The analysis proceeds as follows:

1) spectral modelling

Pre-analysis is done through the Adaptive Lattice algorithm described in ref 4. Successive frames are considered, every 6 ms. The effective length of the time window applied onto the signal is not explicit, but it can be evaluated to about 15 ms. Several maxima, here called "formants" for the sake of simplicity, are usually apparent in each frame.

2) Segmenting the spectral envelope

The formant regions are simply defined between two successive minima of the spectral envelope.

3) filtering in the formant region

The original signal is filtered by short-term Fourier analysis-synthesis (ref 5) in each of the formant regions previously defined. For each frame, N partial signals are therefore obtained; the sum of all of these is equal to the original signal. The chosen filters (rectangular or triangular shape of the transfer function) do not introduce any phase distorsion. The filtering itself is done on a long segment of the original signal (50 ms), surrounding the frame considered, in order not to create any edge effect.

4) temporal envelope peak detection

The temporal envelope of each partial signal is then calculated according to the process described in ref 1; the minimum and maximum values are extracted, and the segment found between two successive minima is processed as the main part of one of the expected wfs, provided that its reference instant (amplitude maximum) appears within the 6 ms frame interval (fig 3). The sum of the detected waveforms is again equal to the partial signal considered. Each elementary waveform represents a local peak in the time-frequency domain.

5) waveform modelling

As shown in II, it appears to be possible to consider the elementary waveforms as the impulse responses of the local filters described above.

We can decompose the F filter into parallel sections. Locally, in proximity

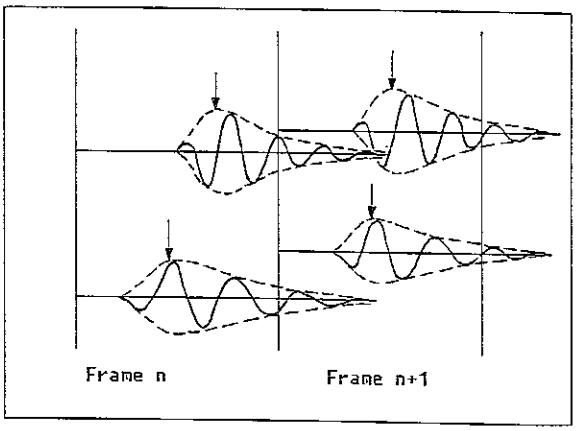


Fig 3 - Waveform behavior at frame boundaries

to a formant, F can be approximated to a second order section having an impulse response which is a sine wave with an exponentially decreasing envelope. If the frequency sectioning function is a rectangular one centered on the formant, the impulse response is modified, but is still able to be modelled in terms of the product of a sine wave and a window.

In order to avoid having an infinitely long waveform, we have introduced an attenuation function, the object of which is to limit effects of overlapping waveforms.

We suppose the following hypotheses:

- the central frequency is constant over the whole length of the waveform.

- the precise form of the envelope is not extremely important.

- an impulse response decreases rapidly. This facilitates its detection by peak-picking of the signal envelope.

The frequency analysis gives the central frequency of each waveform, the temporal analysis gives the reference instant and the time locking of the carrier oscillation, as well as the envelope parameters (amplitude, attack and decay durations).

## IV. Results

The above described system was tested for various male and female voices. Fig 4 shows the beginning of a French sentence uttered by a male speaker, after analysis and evaluation of the main wf parameters. Each detected elementary waveform is represented by a diamond-shaped dot, of heigth proportional to the logarithm of the wf amplitude. Compared to a similar document in ref 1, the grouping of the wfs on the spectral peaks is obviously satis-

factory, in voiced segments as well as in fricative ones. The rapid temporal events (stop releases) also seem to be well represented.

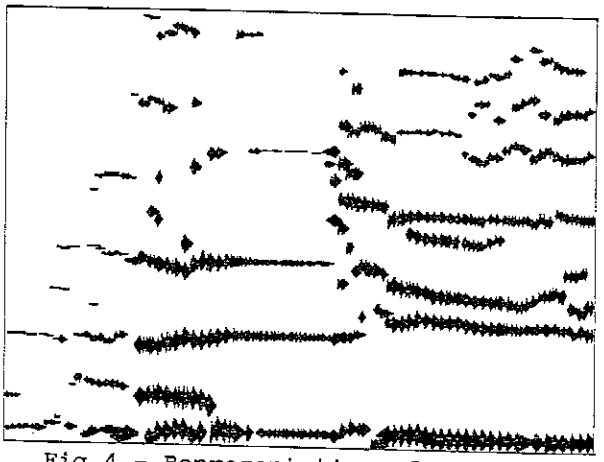

Fig 4 - Representation of a French sentence (beginning of "As-tu vu ce fameux lapin ?", male speaker) as a set of wfm parameters in the time-frequency domain.

From the perceptive point of view, there is no or little loss of quality if the lower part of the spectrum is not taken into account. However, some detection and modelling problems can be found in the part of the spectrum under the F1 region, where parameter estimation errors become perceptually important.

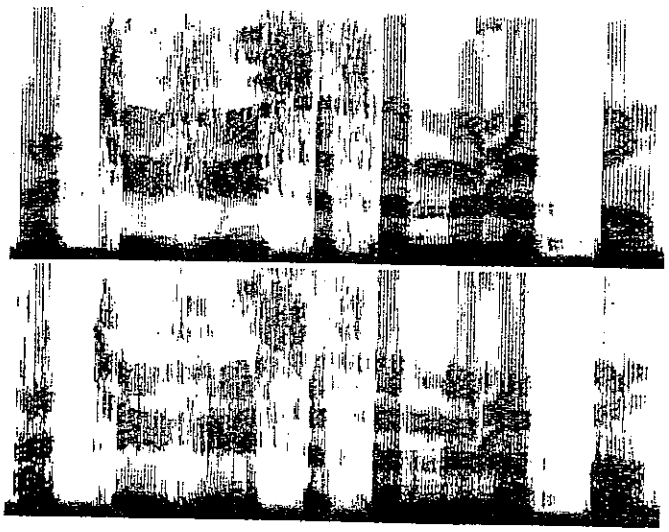

Fig 5 - Conventional spectrograms (top: original; bottom: synthetic) of the entire sentence.

## IV - Conclusion

In this paper we have presented a new manner of analysing the speech signal and of modelling it as a set of elementary waveforms. The goal is the same as in ref 1, i.e. to obtain a description of the signal in terms of entities representative according to the production or perception point of view. Yet the difference lies in the fact that we profit from the structure of the speech signal - poles or spectral maxima - to predetermine the spectral regions where the elementary waveforms should be searched for. The representation we obtain is adequate and can be used as a basis for research into acoustic perceptive structures. Resynthesis yields a signal that is perceptually very close to the original - with the exception of the lowest region of the spectrum (band including Fo) in which some modelling problems remain.

Ref 1 and this paper represent two fairly different points of view; the former is perception-oriented, where the latter uses some characteristics of the speech production system. The premisses are, however, the same. In both cases, as well as in the wavelet approach, a better mastery of the compromise between time and frequency resolutions is sought, and early signal structure decisions are avoided.

## VI - References


1 - Liénard, J.S. "Speech Analysis and Reconstruction Using Short-time, Elementary Waveforms". ICASSP-87, Dallas.

2 - Kronland-Martinet, R.,Morlet, J. and Grossmann, A. "Analysis of Sound Patterns Through Wavelet Transforms". To appear in the International Journal of Pattern Recognition and Artificial Intelligence, special issue on Expert Systems and Pattern Analysis.

3 - Rodet, X. "Time Domain Formant-Wave Function Synthesis", in "Spoken Language Generation and Understanding", J.C.Simon ed., D.Reidel Publishing Co, Dortrecht, Holland, 1980.

4 - Makhoul, J. and Cosell, L. "Adaptive Lattice Analysis of Speech". IEEE Trans. on Circuits and Systems, 28-6, June 1981.

5 - Crochiere, R.E. "A weighted Overlap-Add Method of Short-Time Fourier Analysis-Synthesis", IEEE Trans. on ASSP, 28-1, Feb 1980.