ICASSP 1987
Dallas

SPEECH ANALYSIS AND RECONSTRUCTION USING SHORT-TIME, ELEMENTARY WAVEFORMS

Jean-Sylvain LIENARD

LIMSI-CNRS, Orsay, France
and
Bell Communications Research,
Morristown, New-Jersey

ABSTRACT

We consider the speech signal to be composed of elementary waveforms, wf, (windowed sinusoids), each one defined by a small number of parameters. The typical duration of a wf is of the order of magnitude of a pitch period in the voiced segments, and a few milliseconds in the noise segments. No preliminary evaluation of voicing or pitch is required ; this largely differentiates the approach from the classical pitch-synchronous analysis. The analysis process uses a filterbank, designed to introduce as few time distortions as possible. The signal at the output of each filter is segmented according to successive amplitude minima, and each segment is modeled by a wf. This decomposition can be validated by reconstructing the wfs from their parameters, and summing them in order to recover a signal perceptually equivalent to the original.

I - INTRODUCTION

The speech signal is traditionally analysed with a time lag including several pitch periods, i.e. typically 25 to 50 ms. As a consequence, some sounds like the burst of plosive consonants or the rapid voicing onset of vowels following nasals and fricatives, are eliminated or strongly smoothed, although they may be perceptually relevant. Another feature of the usual analysis methods is that they separate, as early as possible, the properties of the source (voicing, pitch, noise) from the properties of the vocal tract system which is supposed to contain most of the phonetic information.

Another point of view is that the source/-transfer system parameters are not independent from each other; voicing, pitch and noise are actually perceptive qualities of the speech sound, assigned to it at some stage of neural processing, on the basis of the repetition of similar events at regular intervals. In order to investigate this view, we need a spectro-temporal analysis with a time lag shorter than one pitch period, i.e. a few milliseconds; this analysis should be performed before carrying out any interpretation of the signal in terms of Fo, noise, formants, etc.

The present paper deals with a representation of the speech signal using a set of discrete, perceptually relevant, spectro-temporal items called "elementary waveforms" (wf). It is close to the traditional spectrographic representation, which is recognized to be a good basis for the the human (visual) interpretation of the signal. Yet there are two main differences. First, a spectrogram is made of tiny lighter or darker dots or lines that the operator's eye clusters into larger spots, while our process is an attempt to do this grouping automatically. Second, the phase relationship is retained in our process, while it has been lost in the spectrogram, rendering it incapable of a reconstruction of the signal. We propose viewing the speech signal as a set of time-frequency items, "atoms" or "grains", as well as a method for obtaining them. The present paper follows the same line of reasoning as previous work on the short-time analysis of speech (1,2).

The proposed analysis process contains the following steps : filtering with a zero-phase filterbank, and decomposing and modeling the output signals into successive wfs (channel-to-channel modeling). Tentative grouping of the adjacent channels before modeling will also be described.

II - THE FILTERBANK

The filterbank has been designed to meet the requirements of flexibility, computational efficiency, and ability to reconstruct the original signal. Flexibility and efficiency are necessary because our approach is experimental and interactive. To reconstruct the original signal as closely as possible we have to cancel the phase distortions of the individual filters. We use a very simple design (resonator) twice, reversing the time scale during the second pass. The number of filters is user selectable,

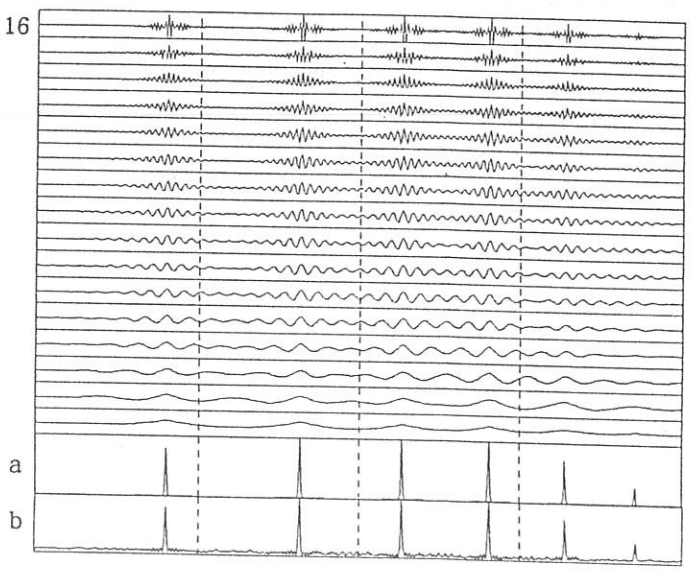

Fig 1 - Decomposition of a series of pulses (a) with a zero-phase filterbank, and signal reconstructed (b) by summation of the 16 output signals. Time scale 10 ms between vertical dotted lines (valid for all the figures in the present paper).

as is their repartition along the frequency scale, which can be linear or non-linear, with a magnification of the lower part of the spectrum. The bandwidths are subjected to a global parameter, which controls their overlap at -6 dB. The gains are adjusted so as to obtain output signals which, after summation over all the channels, produce a good approximation of the original input. By "good approximation" we mean that, with enough filters (practically 10 or 12 in the range 0-5 kHz), there is no perceptible distortion. The reconstructed signal closely ressembles the original (Fig 1).

At the present stage of our study, we do not need to implement a more rigorous solution such as QMF filters, which would yield less flexibility and more computational complexity.

III - THE ELEMENTARY WAVEFORM MODEL

A waveform model (wfm) is a sinusoidal signal multiplied by a windowing function. It is not to be confused with the signal segment, wf, that it is supposed to approximate.Its total duration can be decomposed into attack (before the maximum of the envelope), and decay. In order to minimize spectral ripples, the envelope should present no 1st or 2nd order discontinuity.

X.RODET (3) has defined a wfm which he uses as a component of a parallel-formant synthesizer (Fig 2). The envelope is essentially a decreasing exponential; the damping factor is directly related to the bandwidth of the formant which is to be reproduced. The initial discontinuity is removed through the use of an attack function (raised sinusoid) such that the total envelope is null at the origin, and maximum after a short time. The voiced segments are synthesized by simply adding the 4 or 5 wfms associated with the formants at each pitch period. The noise segments are reconstructed through a more traditional source/filter process. The quality obtained in reproducing a singing voice is extremely good, and the algorithm is computationally efficient.

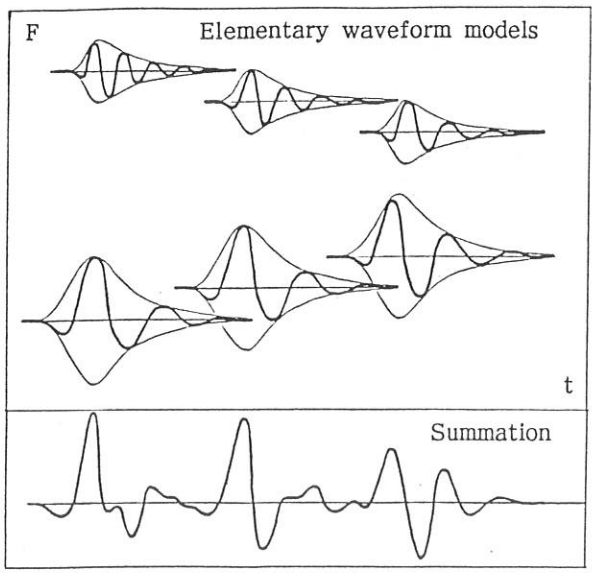

Fig 2 - Parallel-formant synthesis of a voiced segment, by summation of elementary waveform models (principle).

Although exponential damping is natural in the physical world, we choose to model the decaying part of the wfs with another raised sinusoid. Actually we see the wf as a perceptual unit, and not necessarily as the response of a formant filter to a voicing impulse. For example, in noise, the signal goes randomly through energy peaks (energy concentrations in the time-frequency plane) which cannot be thought as systematically having abrupt onsets and exponential decays.

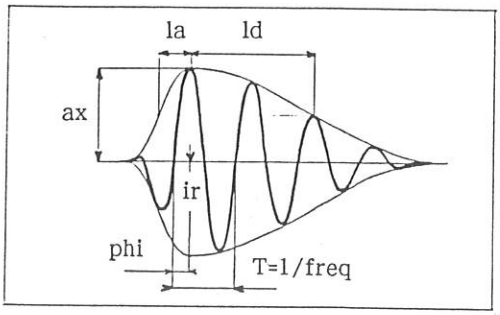

Fig 3 - Waveform model, with attack and decay shaped by raised sinusoids.

Thus a wfm is entirely defined by 6 parameters (Fig 3) : 4 envelope parameters (reference instant ir, attack and decay half-durations la and ld, peak amplitude ax), and 2 carrier parameters (frequency freq=1/T, and phase - or locking delay phi - with respect to the envelope reference instant). This model allows a pure tone to be reproduced by a sum of successive wfms at arbitrary reference instants, which is not possible with an exponential model.

IV - CHANNEL-BY-CHANNEL WF MODELING

Here we suppose that the signal we want to decompose is narrowband in each channel - although for some tunings of the filterbank this hypothesis does not hold, especially in the lowest channels.

The main difficulty is due to the overlap of two or more successive wfs in the signal. The ideal case,

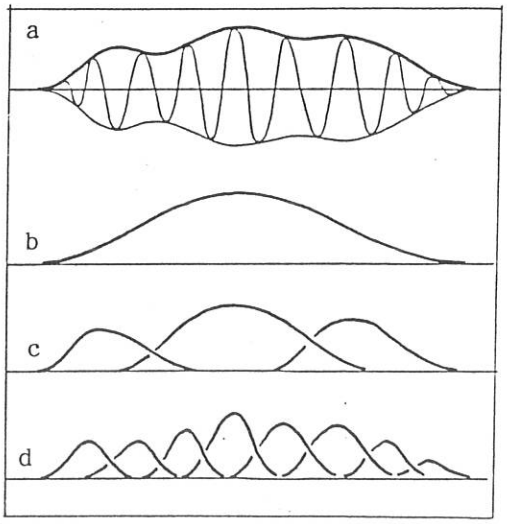

Fig 4 - A signal segment may be approximated by several sets of wfms. Case (b) is to be expected in the high frequency range ; case (d), with one wfm per period, is expected in the lowest channels.

where wfs are well isolated from each other, is not a frequent one. When successive wfs overlap, several decompositions may be possible (Fig 4). In order to obtain a unique decomposition, constraints must be used. These constraints are determined by perceptual experimentation, through analysis and synthesis.

Another difficulty appears in the lowest channels when the instantanous period of the signal comes close to wfm duration. This difficulty is severe when successive wfs overlap.

Our process consists of stating that, to each maximum of the envelope corresponds the maximum of a wfm envelope. It is performed in three steps : evaluating the envelope, smoothing it, and adjusting a wfm to the signal segment surrounding the maximum.

The envelope is determined by full-wave rectification and linear interpolation between successive maxima. This yields a piecewise linear approximation of the envelope. A sampling problem is encountered in the high frequency range, which is dealt with, temporarily, by doubling the sampling rate before rectifying the signal. In the low frequency range, the slight asymmetry between the positive and negative halves of the signal, which usually remains after the filterbank operation, is maintained by the process and used to assign a wfm to each period of the signal.

For each selected reference instant, the "dominant frequency" fd is evaluated over the interval between the surrounding minima. This parameter, computed from the intervals between successive zero-crossings, and weighted by the envelope amplitude, will become the wfm frequency. The phase parameter is computed from the interval between the reference instant and the closest zero-crossing.

The attack and decay parameters are difficult to compute when the wfs overlap. We look for a minimum in the envelope. At that instant the decay of the preceding wfm and the attack of the next are assumed to cross each other at half the amplitude of the envelope. This hypothesis yields a satisfactory reconstruction of the signal when the carrier wave evolves slowly across successive wfs, which tends to be true when the wfs are close to each other. The amplitude parameter is adjusted by comparing the integrals of the wf and of its model, as determined by the former parameters.

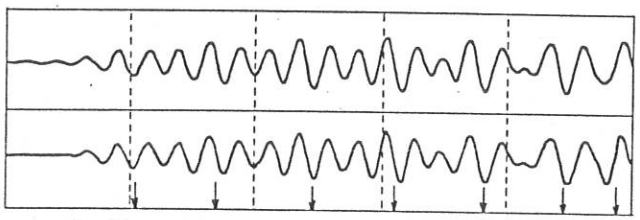

Fig 5 - A narrow-band signal (top) and its reconstruction (bottom) by summation of 7 wfms determined by the modeling process ; the wfm reference instants are marked by arrows pointing downward.

Fig 5 shows that the evaluation process is quite efficient on a narrow-band signal, even with overlapping wfs.

Fig 6 shows a 100 ms segment (vowel /a/ extracted from the beginning of the sentence /atyvysəfamølapɛ̃?/, uttered by a French male speaker), analyzed with a bank of 16 filters having its

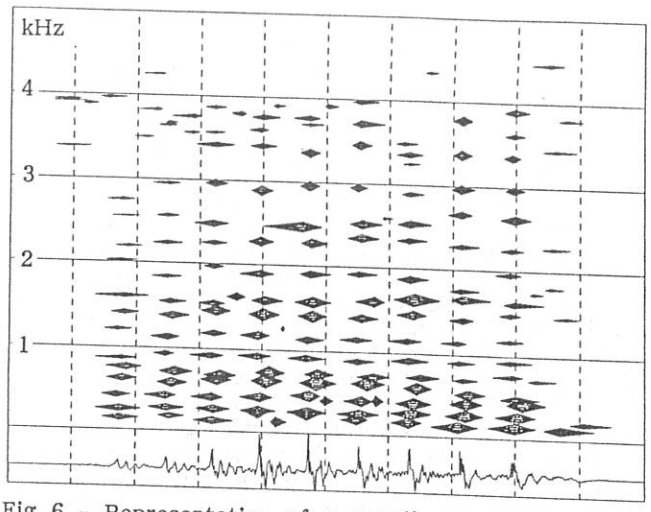

Fig 6 - Representation of a vocalic segment by a set of 170 wfms, displayed as diamond-shaped dots in the time-frequency plane.

center frequencies non-linearly spaced, with bandwidths varying from 125 to 490 Hz at -6 dB. In each channel, the modeling process described above has been applied. The result is a set of 170 wfms, each one defined by its 6 parameters. For an easier visual interpretation, each wfm is represented in the time-frequency plane

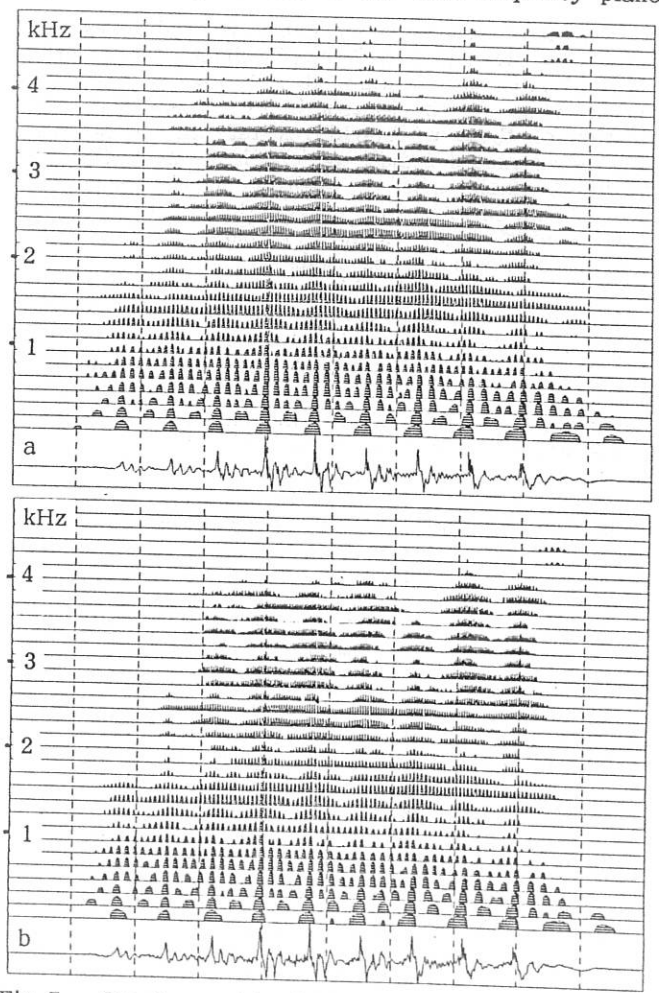

Fig 7 - Spectrographic display of the original (a) and reconstructed (b) vocalic segment of Fig 6.

by a diamond-shaped dot: the height represents the amplitude, the left and right lengths represent the attack and decay durations. The phase parameter is not represented.

Figs 7a and 7b show the original and reconstructed signals (bottom) and their representations in the time-frequency plane. This representation is more precise in time than the usual spectrographic display: the 32 zero-phase filters are linearly spaced; they all have the same bandwidth (200 Hz at -6 dB); in each channel the output signal is displayed after half-wave rectification and logarithmic transformation.

A comparison of the signals themselves, before and after modeling, or of their time-frequency representations shows very few differences. Listening tests, in informal sessions, demonstrate that, for some tunings of the analysis processes, the reconstructed signal can be perceptually almost undistinguishable from the original. Careful listening reveals some attenuation of the high frequency range, as well as a sort of "granularity".

Channel-by-channel modeling can be related to the phase vocoder (4); both use amplitude and instantaneous frequency measurements in frequency channels. They differ in that the phase vocoder smoothes this information, while our process reduces it to a set of discrete elements.

V - MULTI-CHANNEL MODELING

Channel-by-channel modeling is redundant: a given wf in the signal is "viewed" slightly differently by several adjacent filters, and will produce as many different wfms. As the system is totally additive, the summation of those wfms will restitute the proper wf in the reconstructed signal; we would, however, like to associate a single wfm to each wf constituting the original signal, in order to increase the compactness of the representation and to make it independent of the particular filterbank that is used.

One way to obtain this result is to group several adjacent channels in the regions of the time-frequency plane where energy is maximum. The difficulty is in determining which channels are to be grouped, and for how long. We have designed the following iterative process. In each channel the envelope is computed and strongly smoothed, with a zero-phase lowpass filter of cutoff frequency equal to 40 Hz. Between two successive minima the average amplitude is compared to the amplitude of the neighbouring channels, considered within the same time limits. If the observed channel is not a maximum, a fraction of its signal is transfered to the channel of highest amplitude. This is done over all the channels. Some care has to be taken, in order to avoid time discontinuities and errors due to the recursive character of the transfer.

As a result, the list of wfms is much shorter, and the wfms are easier to associate with the usual structures of speech. Comparing Figs 8a and 8b clearly shows that the channel-by-channel wfms have gracefully merged into formant wfms in the voiced segments, into simpler formant-like structures in the fricatives and noise; the burst of the stop consonant is preserved; the pitch appears clearly now as a repetition of similar wfms at regular intervals. The reconstructed speech is perfectly intelligible; it has the right intonation and voicing features. But the quality has decreased. The granularity is more perceptible, and the lowest part of the spectrum has lost some amplitude.

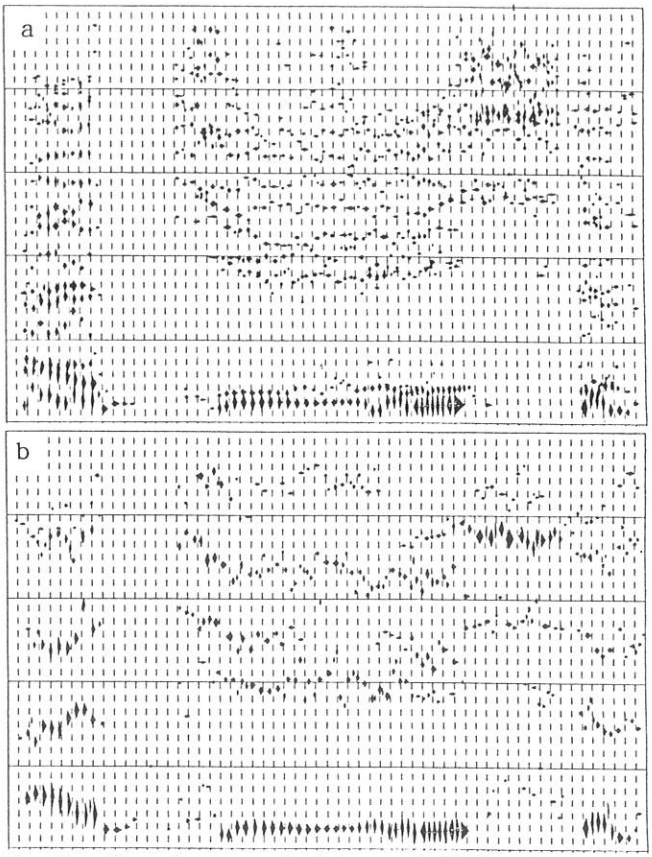

Fig 8 - Speech segment /atyvysə/ displayed as a set of wfms in the time-frequency plane. (a) : channel by channel modeling. (b) : same, after iterative grouping of adjacent channels.

VI - CONCLUSION

A speech signal can be considered as being made of "sound grains", or elementary waveforms defined by a small number of parameters. A decomposition method has been presented, using a zero-phase filterbank analysis, and channel-by-channel modeling of the output signals. A signal which is extremely close to the original can be reconstructed by simply adding the elementary waveforms.

VII - REFERENCES

1 - J.S.LIENARD : "Speech as a String of Pulses ; Pulse-Coherence Function", ASA spring meeting, Ottawa, May 1981.

2 - J.S.LIENARD : "Very Short-Time Analysis of Speech", Notes et Documents du LIMSI, janvier 1985.

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

4 - J.L.FLANAGAN and R.M.GOLDEN : "Phase Vocoder", Bell Syst. Tech. J., vol 45, pp 1493-1509, 1966.

VIII - ACKNOWLEDGEMENTS

This work originated in a preliminary study undertaken during the summer of 1984 at Bell Laboratories, Murray Hill, NJ. The study was pursued and brought to its present state during the author's stay at Bell Communications Research, Morristown, NJ (summer 1986). The author wishes to thank Drs Dan Kahn, Jim Kaiser, Steve Levinson, Aaron Rosenberg and Frank Soong, for helpful discussions and comments on previous drafts on the subject.