TIME-FREQUENCY MODIFICATIONS USING AN ELEMENTARY WAVEFORM SPEECH MODEL

Christophe d'Alessandro
LIMSI-CNRS, BP 133, F-91403 Orsay Cédex, FRANCE

ABSTRACT

An elementary waveform speech model (EWSM) is defined and some capabilities are demonstrated for the modification of localized time-frequency events. The elementary waveforms allow for modelling the local spectro-temporal maxima of energy inside the speech signal by simple mathematical functions. EWSM parameters are estimated using a frame by frame processing: spectral modelling and segmentation using short-time Fourier transform and LPC spectrum, Fourier filtering according to this segmentation, waveforms spotting in each channel waveform modelling with simple functions. The EWSM parameters are relevant according to the classical theory of speech production, and their modifications yield well-localized time-frequency transformations, including frequency compression/expansion, pitch, formant, noise modifications.

1 INTRODUCTION

In this paper we discuss the ability of a new speech signal representation method for time-frequency modifications. *Global* modifications, like frequency expansion/compression, pitch and duration mofications are realized in a very simple way, as well as more unusual *local* modifications allowed by the properties of our representation: such as pitch period, formants, burst, fricative noise modification for instance. The later problem is also treated here in a very simple way, though it remains a difficult task for other representation methods.

We show elsewhere [1] that expansion of the speech signal into a discrete sum of time-frequency well-localized elementary waveforms can be achieved at least from three viewpoints:

- non-parametric methods, short-term Fourier transform and wavelets transform for instance, can receive an elementary waveform interpretation, crossing the classical filterbank analysis and block analysis interpretations [2] [3]. Exact representations and theoretical results are thus available, but some difficulties remain in order to establishing relationship with a speech production or perception model.
- granular analysis [4] based on analogies between auditory models and spectro-temporal analysis. Here, extraction of speech production parameters, formants or pitch for instance is a very interesting problem, but the same kind of difficulties than in auditory modelling are encountered: this point is at present time under study.
- model-based speech elementary waveform decomposition is a continuation of formant waveform synthesis [5]: an elementary waveform speech model (EWSM) can be derived from the classical acoustic model for speech production. The EWSM parameters are thus directly relevant, as speech production parameters. Automatic parameter estimation allows for using this model in the field of speech synthesis or modification.

We will only present and discuss the third approach, for a sake of simplicity, though the first and second ones are able to perform the same type of processing: only interpretation of waveforms parameters from a speech production viewpoint remains more or less difficult in these different cases.

Section 2 introduce the EWSM. Elementary waveforms formulas as well as speech production events viewed trough waveform representation are described.

The automatic analysis/synthesis process, based on spectral segmentation is explained in section 3.

Section 4 deals with the modifications and gives some examples.

In section 5 a conclusion is proposed.

2 ELEMENTARY WAVEFORMS SPEECH MODEL

The EWSM for speech representation is an extension of parallel formant model, in the time domain (figure 1).

The main differences between EWSM and parallel formant model is the lack of excitation/filter distinction in the first case: excitation is only virtual. Thus distinction between source and filter is avoided and the model is clearly located in acoustic domain.

For ideal voiced speech, an elementary formant waveform will be associated to each pitch period, in each formant area. The baseband, defined as the area below the first formant, where the contribution of the glottal airflow waveform is dominant, requires a special treatment: an elementary sinusoïdal parameterization of this contribution is performed.

For ideal unvoiced speech (frication noise), a previous study [6] has experimentaly shown that random generation of elementary waveforms is able, under certain conditions, to produce a noise spectrally equivalent to filtered white noise.

For an actual speech signal, one can easily mix these two ideal cases to produce, for instance, voiced fricatives, stops, or noisy voices.

Thus, two types of elementary waveforms allow for synthesis of both voiced, unvoiced and mixed speech: the next section presents justifications and formulas to choose elementary waveform models.

2.1 formant waveforms

According to the classical acoustic theory of speech production, voiced speech is obtained in the time domain by convolution of an excitation waveform $e(t)$ with the impulse response of a filter $R(t)$ associated to the vocal tract.

$$s(t) = e(t) * R(t) \quad (1)$$

If R is supposed linear and time-invariant, and if excitation is reduced to a train of pulses, parallel decomposition of equation 1 is written in time domain:

$$s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} R_i(t, t_j) \quad (2)$$

where R_i represent the impulse response of the i^{th} parallel section, at time t_j. For a second order section in equation 2, associated with *formants*, the impulse response is:

$$R_i(t) = G_i e^{-\alpha_i t} sin(\omega_i t + \phi_i) \quad (3)$$

where α_i sets bandwidth, G_i amplitude, ω_i central frequency, and ϕ_i phase of the i^{th} formant.

equations 3 and 2 present the behaviour of parallel formant synthesis, with pulse-like excitation in time domain.

We extend this model in two directions: first equation 3 is extended by using a more general formant waveform model proposed by [7], which introduces a smooth attack, and second equation 2 is extended by defining an independant excitation for each formant waveform: it is thus possible to synthesize both periodic and random signals:

$$s(t) = \sum_i G_i \Lambda_i(t) e^{-\alpha_i t} sin(\omega_i t + \phi_i) \quad (4)$$

Λ is a step function, with a cosine rising segment, beginning at reference instant t_i:

$$\Lambda_i(t) = 0 \; for \; t \leq t_i \quad (5)$$

$$\Lambda_i(t) = \frac{A}{2}(1 - cos(\beta(t - t_i))) \; for \; t_i < t \leq t_i + \frac{\pi}{\beta} \quad (6)$$

$$\Lambda_i(t) = 1 \; for \; t > t_i + \frac{\pi}{\beta} \quad (7)$$

equation 4 describes a discrete set of formant waveforms, located at points (t_i, ω_i) in time-frequency plane.

2.2 sinusoïdal parameterization of baseband waveform

For baseband synthesis, using formant waveforms is no more justified, and we propose a short-term sine waveform parameterization , close to [8]. The elementary waveforms are sinusoïdal segments, and the baseband signal is described with a formula close to equation 4:

$$s(t) = \sum_i G_i \Lambda_i(t) sin(\omega_i t + \phi_i) \quad (8)$$

where G_i represents the amplitude, ω_i the frequency, ϕ_i the phase and Λ_i the envelope of the sinusoïdal waveform. Λ_i is a temporal window, made of a rising and a decaying sine for example centered at reference instant t_i.

The complete EWSM combine the two types of waveform, using equations 4 and 8.

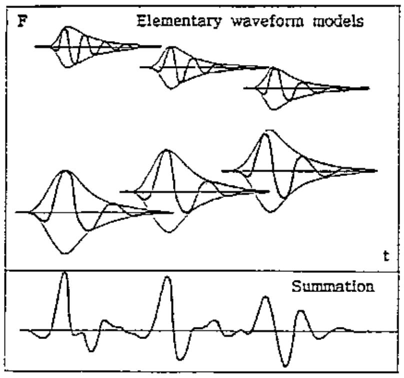

Figure 1: EWSM representation of a voiced segment, by summation of elementary waveforms (principle, from [1] ©IEEE-87)

2.3 EWSM representation of articulatory events

Waveform is the basic element of this representation: thus, an articulatory event is organized as a little set of waveforms.

In the time domain, voiced speech is composed of voicing periods. Each voicing period is composed of formant waveforms sharing a similar reference instant, ideally one in each formant area, and of sine waveforms sharing a similar reference instant, ideally one for each harmonic in the baseband. In frequency domain, voiced speech is composed of formants: a formant is viewed as a set of waveforms sharing similar central frequencies, ideally one waveform for each voicing period. The baseband is decomposed into harmonics: an harmonic is viewed as a set of waveforms sharing similar central frequencies, ideally one waveform for each voicing period.

Unvoiced speech is composed of randomly distributed formant waveforms and sine waveforms, according to the statistics of desired noise (more concentrated in the formant areas, if any).

Burst of stops are composed of a little number of very short-time waveforms, located at the burst instant, and reflecting its spectral composition.

Unvoiced fricatives or noisy voice are obtained by mixing the voiced and the unvoiced case.

For speech modification, the main point is that waveforms parameters are close to production parameters, they represent formant parameters, or voicing period or burst parameters etc., and that each waveform is a basic element which can be treated independently.

3 ANALYSIS/SYNTHESIS PROCESS

An automatic system for EWSM parameter estimation from actual speech has been developed. This system is based on a spectral wideband LPC model in formant area and spectral narrowband STFT model in baseband. Thus spectral local maxima are detected. Spectral segmentation and filtering in these areas give back time domain signals, and temporal segmentation using local temporal maxima allows for detection of natural elementary waveforms. Waveforms parameters are then estimated, and the sum of all synthetic elementary waveforms is the reconstructed signal.

Figure 2 summarize the analysis/synthesis process.

4 TIME-FREQUENCY MODIFICATIONS

The output of the analysis stage, and the input of the synthesis stage, is a set of elementary waveforms described by their parameters. Hence, performing spectro-temporal localized modifications comes to modifying those parameters. This modification is simple to understand, owing to the acoustic relevance of the parameters.

4.1 examples of global modifications

4.1.1 pitch and duration modification

Pitch modification is achieved without explicit pitch extraction. EWSM predict that for ideal voiced speech only one waveform appear for each voicing period. Pitch modification is obtained by modifying only one parameter (the reference instant) for formant waveforms, and by modifying two parameters (the reference instant and the frequency) for sine waveforms. Phase interpolation is achieved by the overlap-add process for sine waveforms. A duration modification occurs with pitch modification. A time domain treatment is used for duration modification alone, wich is not specific to our method [9]. Combining both allows pitch modification without any time distortion.

4.1.2 frequency expansion/compression

Frequency scale expansion/compression is achieved by modification of a single parameter, central frequencies, both for formant and sine waveforms.

4.2 examples of local modifications

Spectro-temporal local modifications of the speech signal are straightforward and simple to understand on the EWSM parameter, provided that the waveforms involved in the modification are well labeled. Thus, the main problem is to assign a set of waveforms to the particular acoustic or articulatory event under study. Automatic waveform labelling is beyond the scope of this paper, and we just attempt to show here the ability of the method for spectro-temporal localized modification. Waveform labelling was manually performed, by visual inspection of EWSM analysis results. Figure 3 is an exemple of such a representation.

4.2.1 formant modification

Amplitude, bandwidth, central frequencie, phase, temporal attack are explicit parameters of the EWSM. Hence, formant modifications are achieved in a very straightforward way. Figure 4.a.b.c gives an example of vowel change. The second formant central frequency is shifted down for all the /α/ to obtain /a/.

4.2.2 noise modification

Modifying the spectro-temporal behaviour of fricative noise is achieved in the same way in time-frequency plane. In figure 4.d.e noise is cut in /s/ and /f/ to obtain /t/ and /p/. In figure 4.f, voicing is drawn out from a /v/, and a little amount of noise is added to obtain a /f/.

5 CONCLUSION

The ability of a new spectro-temporal model-based speech representation for localized modifications has been demonstrated.

Modifications were performed on natural speech trough a high-quality analysis-synthesis system, hence naturalness was preserved.

This method provides a powerfull tool for speech modification, specially suited for phonetic, psychoacoustic and speech synthesis experiments.

REFERENCES

[1] d'Alessandro, C. "Représentation du signal de parole par une somme de fonctions élémentaires". These de doctorat en science, Université Paris VI, April 89 (in french).

[2] Flandrin, P. "Time frequency and time scale". IEEE Fourth Annual ASSP workshop on Spectrum estimation and Modeling, Minneapolis, August 1988.

[3] Combes, J.M., Grossman, A. & Tchamitchian, P. (ed.) "Wavelets, Time-frequency methods and phase space". Springer-Verlag, Berlin, 1989.

[4] Liénard, J.S. "Speech analysis and reconstruction using short-time, elementary waveforms". Proceedings of IEEE-ICASSP-87.

[5] d'Alessandro, C. & Liénard, J.S. "Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation". Proceedings of IEEE-ICASSP 88.

[6] d'Alessandro, C. & Rodet, X. "Synthèse et analyse-synthèse par fonctions d'ondes formantiques". Journal d'Acoustique, No. 2, No.2, June 1989 (in french).

[7] Rodet, X. "Time Domain Formant-Wave-Function Synthesis". in "Spoken Language Generation and Understanding", J.C. Simon ed., D.Reidel publishing compagny, Dordrecht, 1981.

[8] McAulay, R. & Quatieri, T. "Speech Analysis/Synthesis Based on a Sinusoidal Representation". IEEE trans. on ASSP, Vol. ASSP-34, No. 4, 1986.

[9] Neuburg, E.P. "Simple pitch-dependant algorithm for hight-quality speech rate changing." JASA, Vol. 63, No. 2, 1978.

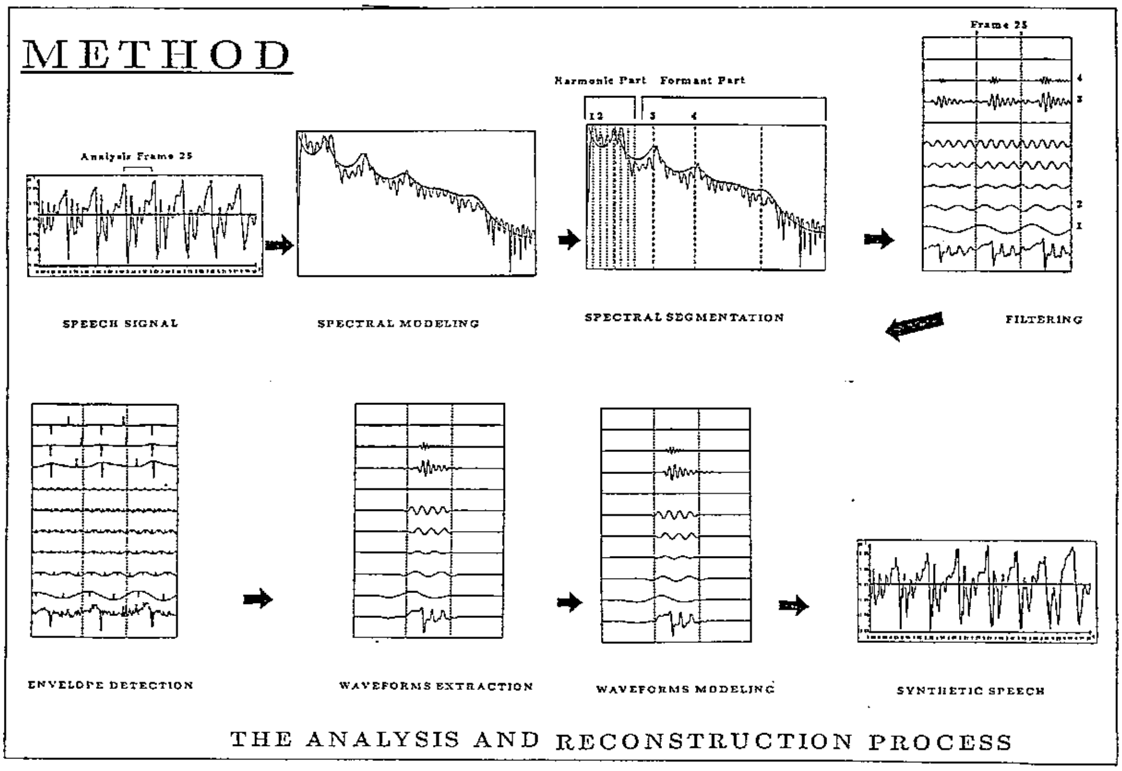

Figure 2: analysis-synthesis process

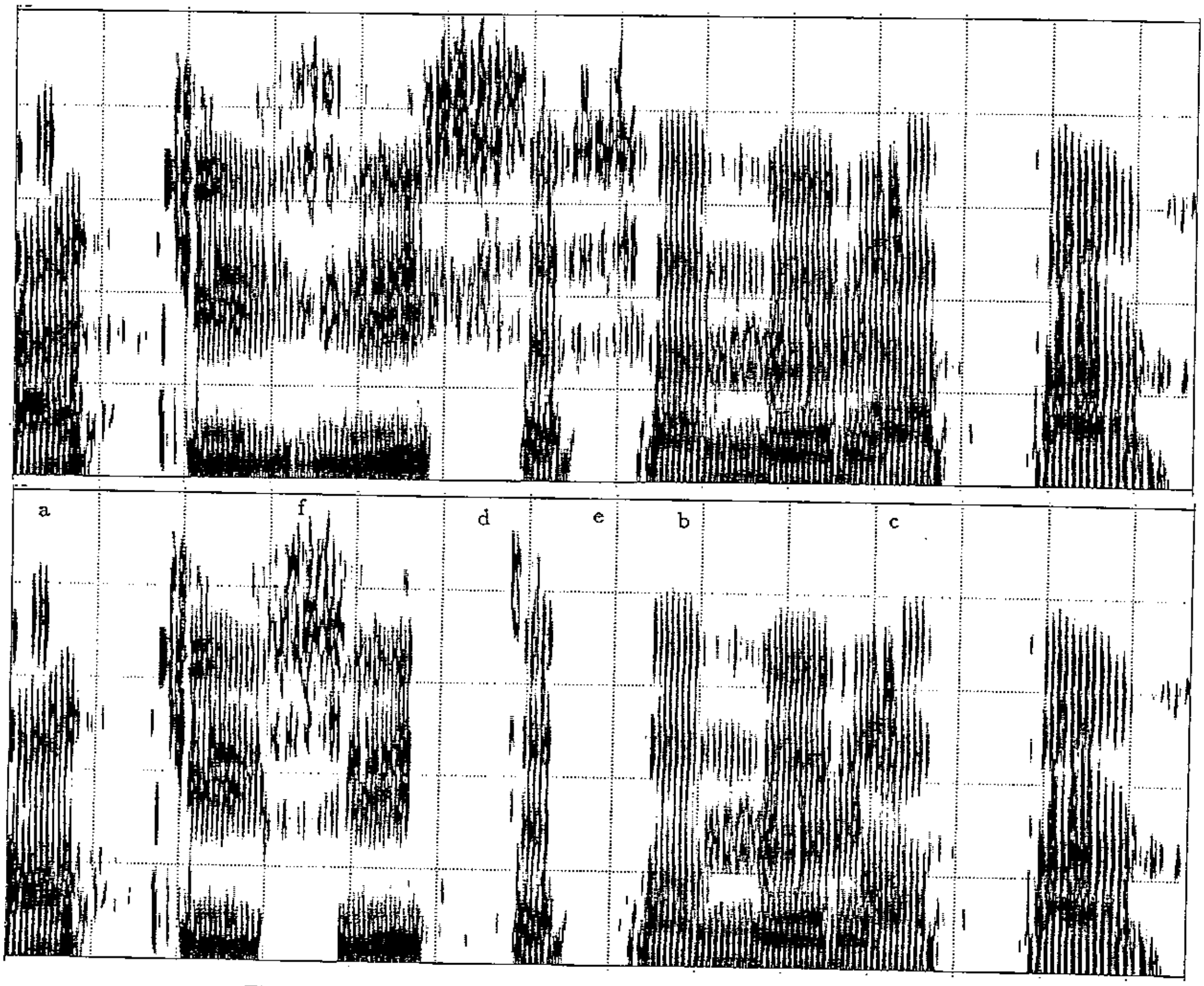

Figure 4: male voice speaking "as tu vu ce fameux lapin ?".
Top: original speech. Bottom: modifyied speech (see text).

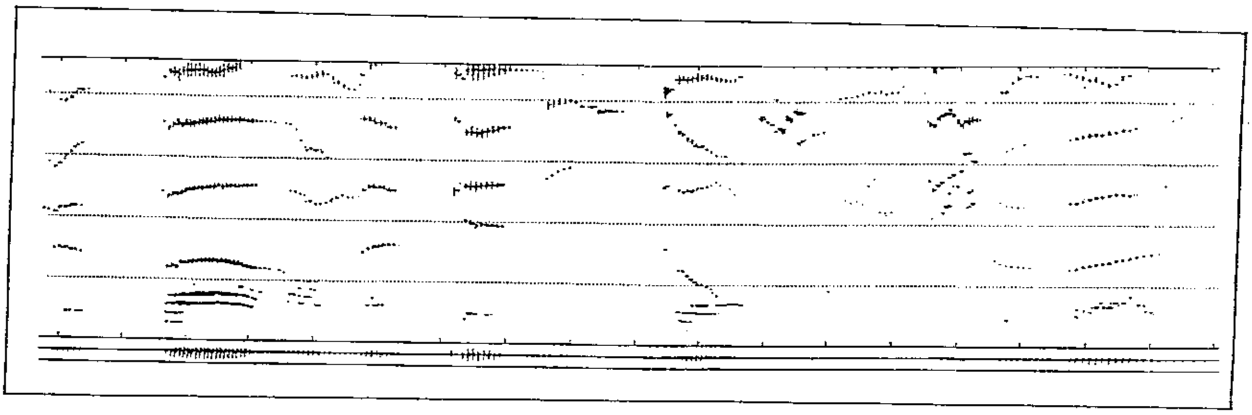

Figure 3: waveforms spotting in the time-frequency plane:
male voice speaking "je vais en Afganistan sur mon cheval".