Speech Communication 9 (1990) 419–431
North-Holland

TIME-FREQUENCY SPEECH TRANSFORMATION BASED ON AN ELEMENTARY WAVEFORM REPRESENTATION

Christophe d'ALESSANDRO
LIMSI-CNRS, BP 133, F-91403 Orsay Cedex, France

Received 27 June 1990
Revised 1 August 1990

Abstract. A representation of the speech signal as a sum of elementary waveforms (Elementary Waveform Speech Model or EWSM) is introduced and some of its features for modifying localized time-frequency events are demonstrated. The elementary waveforms model the local spectro-temporal maxima of energy within the speech signal thanks to the use of simple mathematical functions. An automatic analysis-synthesis system allows for waveforms parameters estimation, using frame-by-frame processing: spectral modelling and segmentation using short-time Fourier transform and LPC spectrum, Fourier filtering according to this segmentation, waveform spotting in each channel, waveform modelling using simple functions. The classical theory of speech production proves the validity of the EWSM parameters: their modifications yield well-localized time-frequency transformations, including frequency compression/expansion, pitch, formant and noise modification.

Zusammenfassung. In diesem Beitrag beschreiben wir eine Darstellung des Sprachsignals als Summe elementarer Funktionen. Diese Darstellung ermöglicht sowohl globale als auch lokale Veränderungen des spektralen und temporalen Inhalts. Wir beginnen mit einer Diskussion der Modellierung des Sprachsignals als Summe elementarer Funktionen und beschreiben anschließend ein automatisches Analyse-Synthesesystem. Die Analyse zur Bestimmung der Modellparameter umfaßt die folgenden Etappen: spektrale Modellierung und Frequenzbandbestimmung durch FFT- und LPC-Analyse, Subbandanalyse mit Hilfe der vorher bestimmten Spektralwerte für jedes Frequenzband, Lokalisierung und Modellierung der elementaren Wellenfunktionen. Die so errechneten Parameter stimmen mit dem linearen akustischen Sprachproduktionsmodell überein und erlauben die Steuerung präziser zeitlicher und spektraler Transformationen wie z.B. Modifizierung des Grundtons, der Formanten oder des Rauschens für Zischlaute.

Résumé. Une représentation du signal de parole comme somme de fonctions élémentaires permet de réaliser simplement des modifications spectro-temporelles, globales ou locales. Après avoir discuté de la modélisation du signal de parole en somme de fonctions élémentaires, un système automatique d'analyse-synthèse est présenté. Les paramètres du modèle sont estimés grâce à une analyse trame par trame: modélisation et segmentation spectrale en utilisant la transformée de Fourier à court terme et la prédiction linéaire, filtrage adaptatif par transformée de Fourier à court terme suivant cette segmentation, détection et modélisation des formes d'ondes élémentaires dans chaque bande d'analyse. Les paramètres du modèle étant pertinents pour le modèle acoustique linéaire de production de la parole, leur manipulation autorise des traitements comme, par exemple, l'expansion/compression spectrale, la modification de la fréquence de voisement, des formants, du bruit de frication.

Keywords. Speech analysis/synthesis, waveform synthesis, formant synthesis, sinusoidal representation, time-frequency modification.

1. Introduction

In this paper we present a new speech signal representation method and an automatic analysis-synthesis system for such time-frequency modifications as frequency expansion/compression, pitch, formant, plosive bursts, and fricative noise modification. It is necessary to be able to perform time-frequency speech modifications in many fields of speech such as acoustic phonetics, synthesis and perception testing. The present study discusses a model and presents the design of an analysis-synthesis system through which natural speech can be locally or globally modified in the

time-frquency domain. *Global* modifications, such as frequency expansion/compression and pitch modifications, as well as less-studied *local* ones (for example, pitch period, formant, burst and fricative noise modification) can be carried out. A complex task for other representation methods, local modification, is dealt with in a very simple way here. These modifications could be useful for investigating the contribution of acoustic parameters to the perception of voice quality, for designing psychoacoustic stimuli from natural speech, for speech representation to be used in formant rule-based or concatenation synthesizers.

Our approach to signal representation for time-frequency modification is to consider speech signal as a discrete set of time-frequency well-localized simple mathematical functions: elementary waveforms (Liénard, 1987; Liénard and d'Alessandro, 1989). We therefore have an event-based decomposition of the speech signal in the time-frequency plane. Each event is centered on a frequency and a point in time, which are locally dominant, and correspond to a local energy concentration. Ideally, the fact that these events correspond to articulatory-phonetic ones means that signal modification is possible, manipulating acoustic-phonetic cues. This correspondence, which is a cornerstone of the method, depends on production-based assumptions for the model's design (Flanagan, 1972). The method does not need an explicit source/filter deconvolution in order to make the manipulations possible: the elementary waveforms and the relationships between elementary waveforms depend on both source and filter. Source/filter deconvolution is valid for describing only part of the acoustic segments which are present in phonation. To give an idea of the relationship between elementary waveform representation and the source/filter concept, a voicing period in a formant area of the spectrum will ideally give birth to one elementary waveform. This elementary waveform is characterized by a series of filter parameters (formant central frequency and bandwidth, which are related to the central frequency and to the time-decay of the waveform), a series of source and filter parameters (formant and voicing amplitudes, which are related to the amplitude of the waveform; spectral slope, which is related to the time-onset of the waveform), and a source parameter (pitch, which is related to the waveform repetition rate).

Two main problems arise in the case of a time-frequency modification system. First, a method for localized time-frequency analysis, modification and reconstruction is needed. Second, in order to interpret these modifications on an acoustic-phonetic basis, analysis and synthesis parameters must be relevant to a production model. To solve the first problem, our approach is to decompose the speech signal in time and frequency, using the short-term Fourier transform: this method, in the filterbank interpretation (Flanagan and Golden, 1966), is useful for adaptive spectral segmentation, and in the block analysis interpretation (Allen, 1977), is useful for time segmentation. For the second problem, we chose to use the classical linear acoustic model for speech production. This model has been used as the basis of much work in acoustic phonetics. Technically, linear prediction analysis and short time Fourier transform analysis can give accurate estimation of this model's parameters. Another, and perhaps better, choice could have been the interpretation of analysis and synthesis parameters in terms of speech perception, or of an auditory model; these possibilities were not pursued due to the limited amount of phonetic knowledge based on auditory processing available at present. Finally, the model and the system presented here represent the speech signal as a sum of elementary waveforms derived from the linear acoustic model of speech production. This representation can be modified, either directly on "natural" (not modelled), "synthetic" (modelled) waveforms, or a mixture of both, depending on the type of modification.

Although we are in a first, exploratory phase of the use of elementary waveforms for speech signal modification, the original ideas here come from three sources. The first is the time-domain elementary waveforms approach for speech synthesis: the Formant-Wave-Function method (Rodet, 1980), used for musical synthesis, and the granular representation of speech (Liénard, 1987) introduced for speech analysis/synthesis. These methods provide the grounds for representing

speech as a sum of elementary waveforms. The second idea concerns the waveform speech modification approach: our method is a generalized form of the cut-and splice method (Scott, 1967; Neuburg, 1978). This technique was introduced for pitch and duration modification some time ago, and interest in it has recently been revived, for speech synthesis (Charpentier and Stella, 1986). The principal idea here is that the speech signal can be cut into small pieces (representing periods of voicing, for example), and that these pieces, called waveforms, can be spliced in a different way to carry out several modifications. Rather than working solely in the time domain, the method presented herein aims at providing time-frequency splicing/modification capabilities. The third original idea concerns parameter estimation using a non-parametric linear signal representation method. Along the lines of the pioneering work of Gabor (1946), we are making use of the time-frequency waveform decomposition provided by the short-term Fourier analysis and synthesis as a basis for our automatic analysis/synthesis system.

This paper is organized in five sections. In Section 2 the Elementary Waveform Speech Model is introduced. Elementary waveform analytic expressions as well as speech production events viewed through the waveform representation are described. The automatic analysis/synthesis process, based on spectral and time domain segmentation is explained in Section 3. Section 4 deals with modifications capabilities and gives some examples. Concluding remarks will be presented in Section 5.

2. Elementary Waveform Speech Model

The Elementary Waveform Speech Model (EWSM) for speech representation is close to the parallel formant model (Flanagan, 1972; Holmes, 1983) and to the sinusoidal model of speech (McAulay and Quatieri, 1986). The main differences between EWSM and the parallel formant model are in the processing of glottal flow and in the time and frequency domain approach in EWSM (a frequency domain approach is used in formant synthesis). In EWSM, no distinction between excitation and filter is made: a harmonic parametrization of the lower part of the spectrum where the glottal flow contribution is dominant is used instead of an explicit glottal flow waveform filtered by formant filters. In parallel formant synthesis the lack of a source/filter distinction appears implicitly in formant areas, due to the formant amplitude control, in spite of the use of a glottal flow waveform. In parallel formant synthesis, formants are considered as frequency domain independent objects, but they are driven by common excitation. Along this line, in EWSM, formants are considered to be independent spectro-temporal acoustic entities. The EWSM is also close to the sinusoidal speech model. Parameters used for the baseband representation are sinusoidal parameters: clearly these parameters are acoustic parameters, not production ones (for example, finding the open quotient of the glottal flow using these parameters is not an easy task), but they are perceptually relevant (for example, changing pitch or loudness, or the amplitude of the second harmonic is quite simple using these parameters). Thus, the EWSM is less realistic from a production standpoint than other terminal-analog synthesis structures (which would give preference to a pole-zero serial structure for the vocal tract, excited by a glottal flow model (Fant, 1960). However, it is somewhat more closely related to perception; it allows for automatic analysis/synthesis and therefore is more attractive for signal manipulation.

Ideally, for 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 flow waveform is dominant, requires special processing. In this case, the model carries out elementary sinusoidal parametrization. On the other hand, for frication noise, a previous study (d'Alessandro, 1989) experimentally showed that random time-frequency generation of elementary waveforms produces a noise spectrally equivalent to filtered white noise. For a real speech signal, one can easily mix these two simple cases to produce, for example, fricatives, stops, or breathy voices. Thus, two types of elementary waveforms permit synthesis of voiced, unvoiced and mixed speech. The next section pre-

sents justifications and analytic expressions for elementary waveform models.

2.1. Formant waveforms

According to the classical acoustic theory of speech production (Fant, 1960), voiced speech is obtained in the time domain by the convolution of a periodic pulse train $\Sigma_i \delta(t - t_i)$, where $\delta(t)$ is the Dirac's distribution, with an excitation waveform $g(t)$, which accounts for the glottal flow waveform, and with the impulse response $v(t)$ of a linear filter modelling the vocal tract.

$$s(t) = \sum_i \delta(t - t_i) * g(t) * v(t). \quad (1)$$

Furthermore, we may define a new filter $r(t)$ accounting for both glottal flow and vocal tract components, as follows:

$$r(t) = g(t) * v(t). \quad (2)$$

Equation (1) reduces to

$$s(t) = \sum_i \delta(t - t_i) * r(t). \quad (3)$$

A parallel decomposition in n sections of the filter, $r(t)$, in equation (3), written in the time domain gives

$$s(t) = \sum_i \sum_{j=1}^{n} r_j(t, t_i), \quad (4)$$

where $r_j(t, t_i)$ represents the impulse of the jth parallel section, excited at time t_i.

If we neglect the glottal flow component in r, we can consider r as a time-invariant second order section, associated to a fixed *formant*, described (for $t \geqslant t_i$) by

$$r_j(t, t_i) = A_j e^{-\alpha_j(t - t_i)} \sin(\omega_j(t - t_i) + \phi_j), \quad (5)$$

where α_j sets the -6 dB magnitude spectrum bandwith ($\alpha_j = \pi B_j$, where B_j is the bandwith), A_j the time-domain amplitude, ω_j the center pulsation ($2\pi f_j$, where f_j is the center frequency), and ϕ_j the initial phase of the jth formant.

Equations (4) and (5) stand for parallel formant synthesis, with pulse-like excitation, in the time-domain.

A more realistic model for $r(t)$ must incorporate a glottal flow component. In the time domain this can be done by introducing a smooth attack for the waveform envelope. Thus, equation (5) can be extended by using a more general formant waveform model $f_j(t, t_i)$, proposed in Rodet (1980), and can be re-stated as follows:

$$f_j(t) = A_j \theta_j(t) e^{-\alpha_j(t - t_i)} \sin(\omega_j(t - t_i) + \phi_j), \quad (6)$$

where, for $t \geqslant 0$, θ_j is a step function, with a raised-cosine segment, culminating at the reference instant $t_i + \pi/\beta_j$:

$$\theta_j(t) = \begin{cases} 0 & \text{if } t \leqslant t_i, \\ \frac{1}{2}(1 - \cos(\beta_j(t - t_i))) & \\ \quad \text{if } t_i < t \leqslant t_i + \pi/\beta_j, & \\ 1 & \text{if } t > t_i + \pi/\beta_j. \end{cases} \quad (7)$$

π/β_j is the envelope attack duration, in the time domain, and controls the width of the "skirts" of the magnitude spectrum of the formant (Rodet, 1980) in the frequency domain without modifying the -6 dB spectral bandwidth.

A simple but efficient way to synthesize unvoiced or mixed speech is to extend equation (4) by defining a separate excitation for each formant: it is thus possible to synthesize both periodic, random and mixed signals. This synthesis scheme is equivalent to a generalized multipulse (Atal and Remde, 1982) synthesis scheme: the transfer function of the system is decomposed according to the formant locations, and independently multipulse-excited in each formant area. The synthesis formula becomes

$$s(t) = \sum_i A_i \theta_i(t) e^{-\alpha_i(t - t_i)} \sin(\omega_i(t - t_i) + \phi_i). \quad (8)$$

Equation (8) describes a discrete set of formant waveforms, located at points (t_i, ω_i) in the time-frequency plane. As found in multipulse linear prediction coding of speech (Atal and Remde, 1982), experimentation shows that an average of 8 to 12 waveforms every 10 milliseconds is necessary and sufficient to produce high quality synthetic speech, and for coding natural speech.

2.2. Baseband and sinusoidal waveforms

For baseband synthesis, the use of formant waveforms is no longer justified; it would create low-frequency quality problems (d'Alessandro and Liénard, 1988). We have therefore adopted a short-terrm sinusoidal waveform parametrization. As in the sinusoidal representation of speech

(McAulay and Quatieri, 1986; Trancoso et al., 1988), the excitation waveform is assumed to be composed of sinusoidal components, having arbitrary amplitudes, frequencies and phases. Nevertheless, two important differences between the sinusoidal model and EWSM remain. First, EWSM limits the use of the sinusoidal parametrization to the baseband only. Second, the concept of sinusoidal component frequency tracks introduced in McAulay and Quatieri (1986) for the sake of frame-to-frame parameter interpolation is no longer useful. As for formant waveforms, sinusoidal waveforms are enveloped sinusoidal segments having a fixed center frequency. Frequency, phase, and amplitude interpolation is achieved by the waveform overlap-add synthesis process. Processing of both voiced and unvoiced speech is possible in this way, as it is in sinusoidal representation. For voiced speech, sinusoidal segments are, of course, related to harmonics. The baseband signal is described by an expression similar to equation (6):

$$s(t) = \sum_i A_i \gamma_i(t) \sin(\omega_i(t - t_i) + \phi_i), \tag{9}$$

where

$$\gamma_i(t) = \begin{cases} 0 & \text{if } t \leqslant t_i, \\ \frac{1}{2}(1 - \cos(\varrho_i(t - t_i))) & \\ \quad \text{if } t_i < t + \pi/\varrho_i, & \\ \frac{1}{2}(1 + \cos(\sigma_i(t - t_i + \pi/\varrho_i))) & \\ \quad \text{if } t_i + \pi/\varrho_i < t < \leqslant t_i + \pi/\varrho_i + \pi/\sigma_i, & \\ 0 & \text{if } t > t_i + \pi/\varrho_i + \pi/\sigma_i, \end{cases} \tag{10}$$

where A_i represents the time-domain amplitude, ω_i the center pulsation ($2\pi f_i$, where f_i is the center frequency), ϕ_i the initial phase, and γ_i the envelope of the sinusoidal waveform. ϕ_i is a temporal window, made of a rising and of a decaying sine (ϱ and σ parameters), centered at the reference instant, $t_i + \pi/\varrho$. Here, the decaying exponential accounting for the formant bandwidth in equation (6) is no longer used. The time domain envelope $\gamma_i(t)$ is now used in order to interpolate the sinusoidal components, and not for spectral shaping, as in formant waveforms.

The complete EWSM combines the two types of waveforms, using equations (6) and (9). The choice of simple time domain functions for elementary waveforms could present problems for at least two reasons: the waveform parameters are invariant during the extent of time covered by the waveform, and the spectral shape of the formant waveform is essentially symmetric. Some quality problems may appear for rapid formant transitions and for precise spectral modelling. Experiments have shown that these potential problems have very little effect on the way the synthesized speech is perceived, especially when both natural and synthetic waveforms are used for speech modifications.

Until now we have discussed an elementary waveform speech *synthesis* model. For speech modification, it is useful to simply segment and modify natural speech. Thus, the EWSM can serve as a basis for a generalized time-frequency *cut-and-splice* (Scott, 1967; Neuburg, 1976) method, performed on natural speech. If we want to modify one formant center frequency, for example, it is necessary to synthesize this formant using the synthetic waveforms described above. But the remaining part of the time-frequency decomposed signal does not need to be synthetic: the natural waveforms (i.e. the natural waveforms extracted with the automatic analysis-synthesis system described below) can be used.

2.3. EWSM representation of articulatory events

The waveform is the basic building block in this representation: an articulatory event is thus organized as a small set of waveforms. A complete study of EWSM representations of articulatory or phonetic events is not within the scope of this paper (but this is future work that must be done in order to implement rule-based speech synthesis using elementary waveforms, for example). Examining the EWSM representations of rough phonetic classes is nevertheless useful to the understanding of its speech modification capabilities. In this paper we will solely look at three fundamental modes of speech production (quasi-periodic excitation, noise, impulse excitation).

In the time domain, voiced speech is composed of periods of voicing. Each period is composed of formant waveforms which share a similar reference instant – there should ideally be one in each formant area and of sine waveforms sharing a

similar reference instant – there should ideally be one for each harmonic in the baseband. In the frequency domain, voiced speech is composed of formants: a formant is viewed as a set of waveforms sharing similar center frequencies – there should ideally be one waveform for each voicing period. The baseband is decomposed into harmonics: a harmonic is viewed as a set of waveforms sharing similar central frequencies – there should ideally be one waveform for each voicing period. Unvoiced speech is composed of randomly distributed formant waveforms and sine waveforms, in compliance with statistics obtained from the noise that the user wishes to synthesize (more concentrated in any eventual formant areas). Stop releases are composed of a small number of waveforms located at the instance of release, reflecting its spectral composition. Unvoiced fricatives and breathy voices are obtained by mixing the voiced and the unvoiced modes.

For time-frequency modification, the main point is that:

(1) the waveform parameters are close to production parameters (they represent formant parameters, voicing or release parameters etc.);
(2) each waveform is a basic element which can be treated independently.

3. Analysis/synthesis process

In this section we discuss the EWSM parameter estimation. An automatic system for EWSM parameter estimation from actual speech has been developed. It represents an utterance as a sum of natural and/or synthetic elementary waveforms. At the first stage, the signal is cut into time-frequency areas, according to a set of spectro-temporal maxima. *Natural* waveforms are obtained at this stage as little pieces of time-windowed frequency-filtered signal. During a second stage these natural waveforms are modelized to obtain *synthetic* waveforms, in accordance with the two waveform models presented above. After these two analysis stages, the synthesis is performed by simply summing the natural and/or synthetic waveforms. The analysis proceeds as follows:

(1) LP spectral modelling. Here, the aim of linear predictive (LP) analysis is to obtain an estimate of the signal magnitude spectrum envelope. To reinforce the high-order spectral maxima, the signal is preemphasized (high-pass first-order recursive filter using a coefficient of 0.95). The LP analysis is carried out by an adaptive lattice filter (Makhoul and Cosell, 1981). This analysis is not pitch synchronous, but uses fixed-length frames, every 6 ms. In the experiments described here, there was a 10 kHz sampling rate, and the order of the model was 12, so that there were at most 6 conjugate pole pairs (associated to spectral maxima) in the 0 to 5000 Hz band. The real length of the time window (exponentially decaying window) applied on the signal is not explicit, but it can be evaluated at about 15 ms. The LP coefficients are Fourier transformed (using a 1024-point FFT), and give a magnitude spectrum estimating the signal magnitude spectrum envelope. Several maxima, called *formants* here, for the sake of simplicity, are usually apparent in each frame.

(2) Segmenting the spectral envelope. The formant regions are defined as being between two successive minima of the spectral envelope. The aim is to roughly segment the magnitude spectrum envelope into a small number of dominant areas (Rodet and Depalle, 1985). These areas should correspond to the actual formant areas for voiced speech. A simple peak-picking algorithm determines the extrema of the spectral envelope. A more refined method (for example LP polynomial root solving) is not necessary for this rough segmentation. For the purpose of conciseness, the term *formant region* will now be employed, although it is clear that these regions are only magnitude spectrum envelope maxima regions, and do not necessarily correspond to actual formants.

(3) Segmenting the baseband. The *baseband* is defined here as the lowest spectral region obtained in the previous stage. In this region, a sinusoidal waveform representation is used. A set of sinusoidal components is calculated, using a Short-Term Fourier Transform (STFT). A narrowband Fourier magnitude spectrum is computed for each frame. A peak-picking algorithm is then applied to the Fourier magnitude spec-

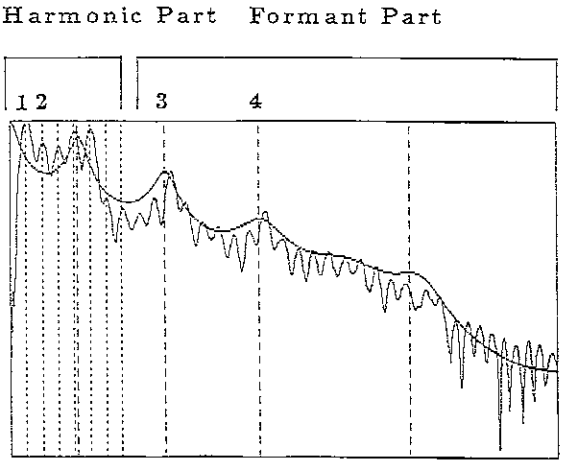

Fig. 1. Spectral modelling of an analysis frame. Spectral segmentation is superimposed on LP and STFT magnitude spectra. (dotted lines: harmonic maxima; continuous lines: formant maxima; frequency range 5 kHz). Nos. 1, 2 represent two harmonic regions. Nos. 2, 4 represent two formant regions.

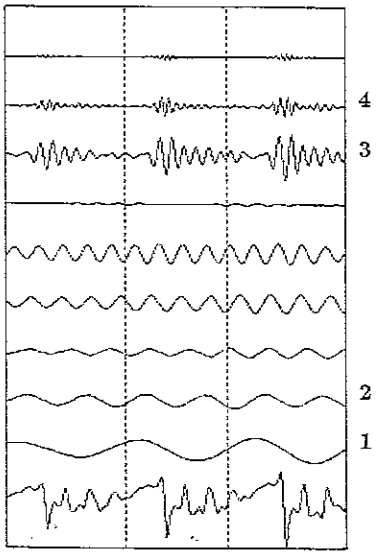

Fig. 2. Filtered signals. The dotted lines represent the limits of an analysis frame (6 ms). The first signal is the reconstructed signal (sum of the filtered signals). The filtered signals corresponding to the harmonic regions 1, 2 and to the formant regions 3, 4 of Figure 1 have the same No.

trum, to obtain a set of maxima. These maxima are expected to correspond to harmonics in voiced speech. The term *harmonic regions* will now be employed, although it is clear that these regions are only narrow-band STFT magnitude spectrum maxima regions, and do not necessarily correspond to actual harmonics.

(4) Filtering in formant and harmonic regions. The original signal is filtered in each of the previously defined formant and harmonic regions. STFT analysis-synthesis is used to perform this adaptive filtering. A 20 ms Hamming window is first applied to the signal. STFT is performed using a 256-point FFT every 5 ms (overlap factor of 4 in the time domain). Since linear time-invariant filtering is being implemented, no synthesis window is needed, and the STFT synthesis is performed with the weighted overlap-add method (Crochière, 1980). For each frame, N partial signals are therefore obtained, the sum of all of these being equal to the original signal. The filter gains are ideally rectangular (in fact they are products of the convolution of the rectangular spectral windows by the Hamming window's spectrum). These filters do not introduce any phase distortion. The filtering itself is done on a long segment of the original signal (50 ms), surrounding the frame being processed, in order not to create any edge effects.

(5) Formant waveforms: temporal envelope peak detection. Above the baseband the following processing is used. The temporal envelope of each partial signal is calculated according to the process described in (Liénard, 1987). The partial signal is full-wave rectified. Then the maxima of this signal are linked using piecewise affine functions. The new signal is low-pass filtered by a zero-phase filter to obtain a smooth temporal envelope function. The minimum and maximum values are extracted, and the segment that is between two successive minima is processed as the main part of one of the expected waveforms, provided that its reference instant (amplitude maximum) appears within the 6 ms frame interval. The sum of the detected *natural* waveforms is again equal to the given partial signal. Each elementary waveform represents a local peak in the time-frequency domain.

(6) Formant waveform modelling. As shown in the previous section, it appears to be possible to consider the elementary waveforms in the formant regions as the responses of the local filters decribed above. For each waveform, six parameters must be evaluated: temporal amplitude A_i, center frequency f_i, initial phase ϕ_i, bandwith B_i (or α_i temporal decay-rate), attack duration π/β_i, and the reference instant t_i. The center frequency f_i is evaluated in the frequency domain, using the magnitude spectrum envelope maxima. Other estimates of this parameter are the zero-crossing rate of the filtered signal, and a smoothed version of the instantaneous frequency. The first solution was used due to the accuracy of its estimates. The five other parameters are estimated in the time domain. The reference instant is determined through the temporal envelope maxima. The first zero-crossing after this instant is used to estimate the phase of the waveform. The envelope parameters (attack duration, and decay rate linked to bandwidth) are set according to the minima of the envelope on both sides of the maxima. The temporal amplitude of the waveform is calculated to equate energies of the natural and synthetic waveforms.

(7) Harmonic waveform modelling. The baseband requires special processing: the temporal envelope in harmonic regions is almost flat. No envelope computation is needed here, owing to the narrowband nature of these signals. Temporal segmentation is attained by extrema detection directly on the filtered signals. Six parameters must be estimated for this type of waveform as well: temporal amplitude A_i, center frequency f_i, initial phase ϕ_i, the two envelope parameters ϱ_i and σ_i, and the reference instant t_i. Parameter estimation is carried out in the same way as it is for formant waveforms. On the other hand, temporal segmentation in baseband is less meaningful than in the formant regions, since signals in harmonic regions are weakly amplitude modulated. In formant regions, amplitude modulation of the filtered signals is linked to the excitation: for voiced speech, pitch periods consequently produce amplitude modulation. In the harmonic region this phenomenon is not present, thus temporal segmentation appears somewhat arbitrary,

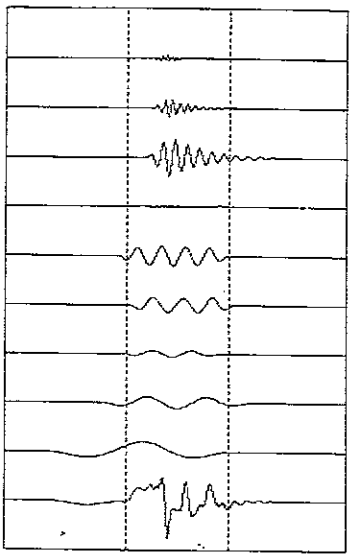

Fig. 3. Waveform modelling. The synthetic waveforms (to be compared with the filtered signals of Figure 2) of an analysis frame are shown. The first piece of signal is the reconstructed signal, summing these waveforms.

from an acoustic standpoint, and is introduced for the sake of interpolation and modifications.

According to the waveform models that are used, some hypotheses must be put forward about the nature of the signal: spectral segmentation is relevant (i.e. the signal contains spectral maxima); the EWSM parameters are invariant over the whole length of the waveform; the precise time domain envelope of the waveform is not extremely important, and can be approximated with simple models; in formant regions an impulse response decreases rapidly, facilitating its detection by peak-picking on the signal envelope.

The above-described system was tested for various male and female voices. Figure 1 shows the LP and STFT spectral modelling (Magnitude spectra), and the result of the spectral segmentation process. Figure 2 shows the filtered signal in the maxima regions defined at the previous stages. Figure 3 shows the modelled (synthetic) waveforms extracted in one analysis frame. Figure 4 shows the French segment /tiʃa/, after analysis. The filtered signals (reconstructed using the natural waveforms) in each formant or har-

monic region are drawn, centered at their center frequencies. Using natural waveforms, the signal is obviously exactly reconstructed. Using synthetic waveforms, there is very little loss of quality, if any.

3. Time-frequency modifications

The output of the analysis stage, and the input of the synthesis stage, are a set of elementary waveforms described by their parameters (synthetic waveforms) and by their samples (natural waveforms). Hence, time domain and/or frequency domain modifications changes these parameters. These modifications are easy to understand, owing to the acoustic relevance of the parameters. The method gives a unified framework for various modifications, but is probably not better than other *ad hoc* methods for global modification. Consequently, we will more fully discuss the localized time-frequency modifications for which this method appears particularly well-suited.

4.1. Examples of local modifications

Spectro-temporal local modifications of the speech signal are straightforward and simple to understand using the EWSM parameter, provided that the waveforms involved in the modification are well-marked. 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; this paper will simply show the usefulness of the method for spectro-temporal localized modification. Waveforms can easily be marked using their time-frequency coordinates (center frequency and reference instant). Some examples of localized time-frequency modifications follow, this list is practically limitless.

4.1.1. Formant modifications

For voiced speech, at least, formants are essential entities to the field of phonetic acoustics. Modifying the formant parameters is therefore an important problem for many branches of speech sciences (Kuwabara, 1984; Makhoul, 1976).

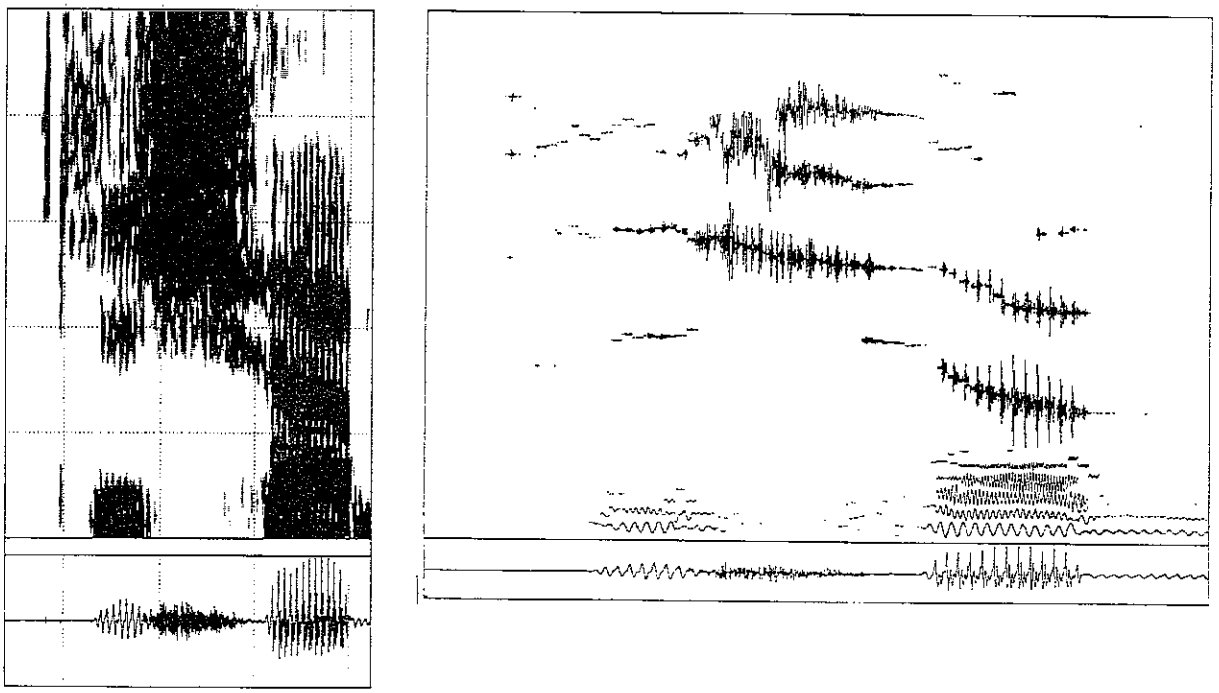

Fig. 4. Representation of the waveforms in the time-frequency domain (right). The same (/tiʃa/) segment is shown in wideband spectrographic format (left: duration 0.55 s: frequency range: 5 kHz).

Mathematically speaking a formant is defined as a maximum of the transfer function of the filter associated with the vocal tract: thus a formant is linked to a conjugate pair of poles, and described only in the frequency domain by the arguments of the poles (which are opposite and which give the formant center frequency) and their modulus (which are equal, and give the formant bandwidth). Since they are considered more as phonetic entities than as mathematical entities, formants have a somewhat richer description, which does not correspond to formant parameters, strictly speaking, rather to formant parameters as they may be observed in a spectrogram, or as they are needed in formant speech synthesis. Therefore formants are usually described by their spectral parameters: center frequencies, bandwidth, spectral amplitudes. These parameters are explicit parameters of the EWSM. Temporal parameters are also available here: initial phase, reference instant (linked to the excitation instant). The attack duration, as mentioned above, is another parameter useful for precisely controlling the time domain (onset time) and the correlative frequency domain (spectral skirts) behavior of the formant. Formant frequency, bandwidth and amplitude modifications are carried out by a simple multiplication of the corresponding parameter. Phase, attack duration and above all reference instant are more unusual formant parameters but may also be easily modified.

Somewhat different processing is needed for the first spectral maximum. This maximum can be (or not) the first formant combined with the source's spectral contribution. Modification of the maximum location, bandwidth or amplitude is carried out in the frequency domain through harmonic waveform amplitude modifications. Formant phase modifications in the baseband require the modification of the phase pattern of the set of sinusoidal components.

Figure 5 gives an example of changing one vowel to another. The second formant central frequency has been shifted down.

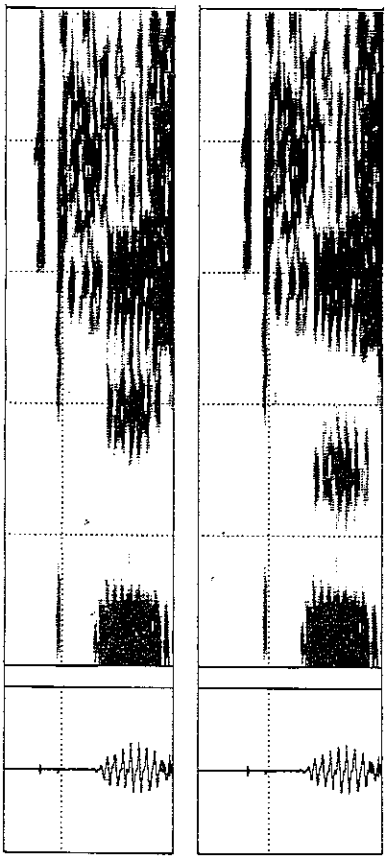

Fig. 5. Local modification: the second formant of the vowel is shifted down (left: natural; right modified: duration 0.15 s; frequency range 5 kHz).

4.1.2. Fricative noise modification

For modifications, fricative noise parameters have not been studied as thoroughly, and so are less well-understood than formant parameters. Noise is viewed as a random distribution of waveforms in the EWSM representation. Each noise spectro-temporal event (or noise *grain*, again using the terminology of Liénard and d'Alessandro (1989)) is found by its spectro-temporal coordinates, and can be processed through these coordinates. The other waveform parameters also vary randomly, and are not easy to interpret except on an average. The granular image appears well-suited for performing noise modification, and thus modification of the spectro-temporal behavior of fricative noise is carried out similarly to formant modifications in the time-frequency plane.

In Figure 6, a/ʃ/ is cut in order to create a /k/ burst.

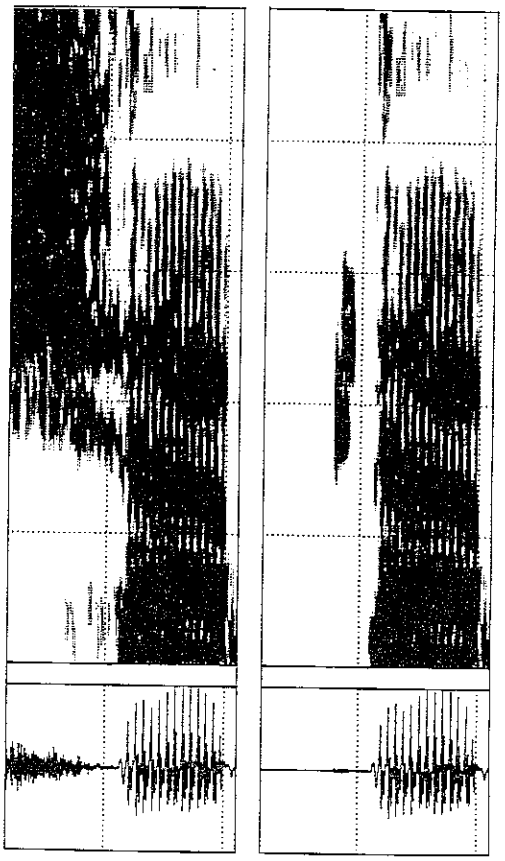

Fig. 6. Local modification: fricative noise of a /ʃ/ is cut in time and frequency to create a /k/ (left: natural; right modified: duration 0.19 s; frequency range 5 kHz).

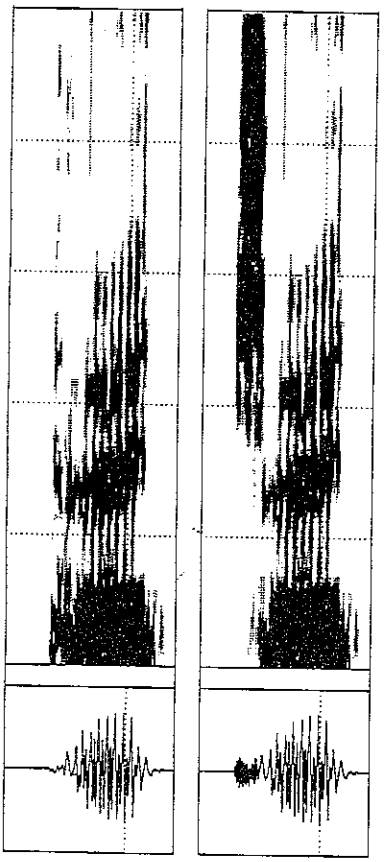

Fig. 7. Local modification: some noise (extracted from a /s/) is added to the burst of a /p/ to create ... a phonetic monster (left: natural; right modified: duration 0.12 s; frequency range 5 kHz).

4.1.3. *Plosive burst modification*

Another example of localized time-frequency modification is plosive burst modification. Bursts appear as a small set of waveforms, in both the formant and harmonic regions. Modifications on both the spectral and temporal aspects of these acoustic events are easy. The EWSM allows the burst components to be moved independently of one another in the time-frequency domain; their individual frequency and time extents can be changed, for example. As in the case of the previous types of modifications, natural waveforms can be used everywhere except where the modification is to take place.

In Figure 7 noise (from an /s/) was added to the burst of a /p/.

4.2. *Examples of global modifications*

4.2.1. *Pitch modification*

Pitch modification (Quatieri and McAulay, 1986; Seneff, 1982; Portnoff, 1981) does not necessitate explicit pitch extraction (using a generalized cut-and-splice method). Pitch periods are implicitly present in the formant and harmonic waveform parameters. The EWSM predict that for ideal voiced speech at least one, or an integral number of waveforms appear for each voicing period in each formant or harmonic region.

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 carried out by the overlap-add process for sine waveforms. Modelling is not needed for pitch modification for formant waveforms: modifying the reference instant of a natural waveform is carried out by shifting its samples before summation. For harmonic waveforms, modification of the reference instant is not sufficient because pitch is involved both in the repetition frequency (linked to reference instants, as for formant waveforms) and in the center frequencies of the waveforms. It is therefore a two-stage modification. First, new synthetic waveforms are obtained by center frequency modifications (multiplication of the center frequencies by a constant factor corresponding to the pitch modification), preserving the same reference instant and the same initial phase. Second, the reference instants are shifted according to the inverse factor that was used for center frequency modifications. Voiced and unvoiced speech are processed in the same manner. This method is derived from the pitch period cut-and-splice method, and the quality obtained is comparable. Our method is, of course, much more complicated and costly in terms of computation time than the time domain processing of the cut-and-splice method, but no pitch extraction is necessary.

Duration modification occurs in conjunction with the pitch modification. Time domain processing can be used for modifying duration alone, (this is not specific to our method (Charpentier and Stella, 1986)). Combining both pitch and duration modifications gives a pitch modification that has no time distortion.

4.2.2. Frequency expansion/compression

Frequency scale expansion/compression and pitch modification are symmetric operations, from the production standpoint (Quatieri and McAulay, 1986; Malah, 1979): in frequency expansion/compression, the vocal tract filter is modified and the excitation remains the same, although in pitch modification the excitation is modified and the filter remains invariant. In the formant region, frequency expansion/compression is carried out by multiplication of the formant waveform center frequencies. Bandwidth modification can be done independently. For harmonic waveforms, frequency compression/expansion requires modification of the amplitude parameters only: the spectral amplitudes of each component must be adapted according to the frequency scale alteration.

5. Conclusion

This paper has presented a new manner of analysing the speech signal and of modelling it as a set of elementary waveforms. The goal is to obtain a description of the signal in terms of entities that are representative according to the acoustic speech production model. The structure of the speech signal is used to predetermine the spectral regions where the elementary waveforms should have been searched for. The baseband, where a formant-based formalism was no longer adequate, required refined modelling. We used an elementary waveform sinusoidal representation for this spectral region. As the waveform parameters were chosen to be relevant from an acoustic point of view, a wide range of speech modifications have easily been performed by modifying these parameters. The powerful capabilities of this new spectro-temporal model-based speech representation for localized modifications have been demonstrated. Modifications were performed on natural speech through a high-quality automatic analysis-synthesis system, hence naturalness was preserved. This method provides an essential tool for speech modification, especially suited for phonetic, psychoacoustic and speech synthesis experiments.

This work on the elementary waveform representation and application to speech modification and synthesis could be extended in two directions. First, a systematic study of speech event representation as a sum of elementary waveforms is needed, both for speech synthesis and for speech modification. Automatic marking of the waveforms according to the phonetic context, and grouping into structured sets of waveforms would also be of great use. Second, we must extend the perceptive aspect of the representation. This aspect is already present in the parameter choices, and in the localized time-frequency decomposi-

tion, but must be studied thoroughly. An auditory-based frequency scale, and information on and use of auditory grouping techniques from what is known so far, should furnish interesting new directions for the evolution of this work.

Acknowledgments

The author would like to thank Jean-Sylvain Liénard and Xavier Rodet for their suggestions and encouragement during this research, Jean-Sylvain Liénard, Maxine Eskénazi and Maria-Gabriella Di Benedetto for helpful comments on earlier versions of the paper, and Martine Adda-Decker for help with the German abstract.

References

C. d'Alessandro (1989), *Représentation du signal de parole par une somme de fonctions élémentaires*, Thèse de doctorat, Université Paris VI, Ch. 3.4, pp. 135–151 (in French).

C. d'Alessandro and J.S. Liénard (1988), "Decomposition of the speech signal into short-time waveforms using spectral segmentation", *Proc. IEEE, ICASSP-88, New York*, pp. 351–354.

J.B. Allen (1977), "Short-term spectral analysis synthesis, and modification by discrete Fourier transform", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP 25, No. 3, pp. 235–238.

B.S. Atal and J.R. Remde (1982), "A new model of LPC excitation for producing natural-sounding speech at low bit rates", *Proc. IEEE, ICASSP-82, Paris*, pp. 614–617.

F.J. Charpentier and M.G. Stella (1986), "Diphone synthesis using an overlap-add technique for speech waveforms concatenation", *Proc. IEEE, ICASSP-86, Tokyo*, pp. 2015–2018.

R.E. Crochiere (1980), "A weighted overlap-add method of short-time Fourier analysis/synthesis", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP 28, No. 1, pp. 99–102.

G. Fant (1960), *The Acoustic Theory of Speech Production* (Mouton, The Hague), pp. 15–90.

J.L. Flanagan (1972), *Speech Analysis, Synthesis and Perception* (Springer, Berlin), pp. 204–272.

J.L. Flanagan and R. Golden (1966), "Phase vocoder", *Bell Syst. Tech. J.*, Vol. 45, pp. 1493–1509.

D. Gabor (1946), "Theory of communication", *J. IEE*, No. 93, pp. 429–457.

J. Holmes (1983), "Research report: Formant synthesizer: Cascade or parallel", *Speech Comm.*, Vol. 2, No. 4, pp. 251–273.

H. Kuwabara (1984), "A pitch-synchronous analysis/synthesis system to independently modify formant frequencies and bandwidth for voiced speech", *Speech Comm.*, Vol. 3, pp. 211–220.

J.S. Liénard (1987), "Speech analysis and reconstruction using short-time, elementary waveforms", *Proc. IEEE ICASSP-87, Dallas*, pp. 948–951.

J.S. Liénard and C. d'Alessandro (1989), "Wavelets and granular analysis of speech", in: *Wavelets, Time-Frequency Methods and Phase Space*, ed. by J.M. Combes, A. Grossmann and Ph. Tchmitchian (Springer, Berlin), pp. 158–163.

R.J. McAulay and T.F. Quatieri (1986), "Speech analysis/synthesis based on a sinusoidal representation", *IEEE TRrans. Acoust., Speech, Signal Proc.*, Vol. ASSP 34, No. 4, pp. 744–754.

D. Malah (1979), "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signal", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP 27, No. 2, pp. 121–133.

J. Makhoul (1976), "Methods for nonlinear spectral distortion of speech signals", *Proc. IEEE ICASSP-76, Philadelphia*, pp. 87–90.

J. Makhoul and L. Cosell (1981), "Adaptive lattice analysis of speech", *IEEE Trans. Circuits and Systems*, Vol. CAS 28, Vol. 6, pp. 494–498.

E.P. Neuburg (1978), "Simple pitch-dependent algorithm for high quality speech rate changing", *J. Acoust. Soc. Am.*, Vol. 63, No. 2, pp. 624–625.

M.R. Portnoff (1981), "Time-scale modification of speech based on short-time Fourier analysis", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP 29, No. 3, pp. 374–390.

T.F. Quatieri and R.J. McAulay (1986), "Speech transformation based on a sinusoidal representation", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP 34, No. 6, pp. 1449–1464.

X. Rodet (1980), "Time domain formant-wave-function synthesis", in: *Spoken Language Generation and Understanding*, ed. by J.C. Simon (Reidel, Dordrecht).

X. Rodet and P. Depalle (1985), "Synthesis by rules: LPC diphones and calculation of formants trajectories", *Proc. IEEE ICASSP-85, Tampa*, pp. 736–739.

R.J. Scott (1967), "Time adjustment in speech synthesis", *J. Acoust. Soc. Am.*, Vol. 41, pp. 60–65.

S.S. Seneff (1982), "Speech transformation without pitch extraction", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP 30, No. 4, pp. 566–578.

I.M. Trancoso, L.B. Almeida, J.S. Rodrigues, J.S. Marques and J.M. Tribolet (1988), "Harmonic coding – State of the art and future trends", *Speech Comm.*, Vol. 7, pp. 239–245.