ELSEVIER

Speech Communication 19 (1996) 221–244

SPEECH COMMUNICATION

Analysis/synthesis and modification of the speech aperiodic component [1]

Gaël Richard [a,b,*], Christophe d'Alessandro [a,2]

[a] *LIMSI-CNRS, BP 133, F-91403 Orsay Cedex, France*

[b] *CAIP Center, Rutgers University, Frelinghuysen Road, CN 1390, Piscataway, NJ 08855-1390, USA*

Received 13 September 1994; revised 2 July 1996

Abstract

The general framework of this paper is speech analysis and synthesis. The speech signal may be separated into two components: (1) a periodic component (which includes the quasi-periodic or voiced sounds produced by regular vocal cord vibrations); (2) an aperiodic component (which includes the non-periodic part of voiced sounds (e.g. fricative noise in /v/) or sound emitted without any vocal cord vibration (e.g. unvoiced fricatives, or plosives)). This work is intended to contribute to a precise modelling of this second component and particularly of modulated noises. Firstly, a synthesis method, inspired by the "shot noise effect", is introduced. This technique uses random point processes which define the times of arrival of spectral events (represented by Formant Wave Form (FWF)). Based on the theoretical framework provided by the Rice representation and the random modulation theory, an analysis/synthesis scheme is proposed. Perception tests show that this method allows to synthesize very natural speech signals. The representation proposed also brings new types of voice quality modifications (time scaling, vocal effort, breathiness of a voice, etc.).

Zusammenfassung

Dieser Artikel behandelt maschinelle Sprachsynthese. Das Sprachsignal kann in zwei Hauptbestandteile eingeteilt werden: (1) die periodische Komponente enthält quasi-periodische oder stimmhafte Anteile und wird durch quasi-reguläre Stimmbandvibrationen der Stimmbänder erzeugt; (2) die aperiodische oder Rauschkomponente ist rein zufälliger Natur. Sie entsteht bei stimmhaften Lauten, wie z.B. den frikativen anteilen im Phonem /v/, oder in Abwesenheit von Stimmbandvibrationen, wie z.B. bei den Frikativen /s/, /t/, etc. Diese Arbeit soll zur exakten Modelierung jener zweiten Komponente und insbesondere der Signale des modulierten Rauschens, dem Rauschen der stimmhaften Frikative, beitragen. "Shot-noise-Effekte" begründen die im ersten Teil eingeführte Synthese-Methode. Die im Rahmen dieser Technik verwendeten Punktprozesse definieren Ankunftsmomente spektraler Ereignisse, dargestellt durch Formantierte Wellenformen (FWF). Es folgt ein theoretischer Ansatz, der sich auf die Darstellung nach Rice stützt. Vorgestellt wird ein Algorithmus zur Analyse und Synthese jener aperiodischen Komponente im Zeitbereich. Perzeptionstests haben gezeigt, daß die im Rahmen dieser

* Corresponding author. E-mail: gael@caip.rutgers.edu.

[1] Audiofiles available. See http://www.elsevier.nl/locate/specom .

[2] E-mail: cda@limsi.fr.

PII S0167-6393(96)00038-6

Technik erzeugte Sprache als natürlich empfunden wird. Der vorgestellte Ansatz eröffnet außerdem neue Perspektiven zur Modifikation der stimmqualität (Veränderungen zeitlicher Natur, Sprachintensität, Artikulation, etc.).

Résumé

Le cadre général de ce travail est celui de l'analyse et de la synthèse de la parole par ordinateur. Le signal de parole peut être scindé en deux composantes principales: (1) une composante périodique (constituée des éléments quasi-périodiques (ou voisés) produits par une vibration quasi-régulière des cordes vocales); (2) une composante apériodique ou de bruit (constituée des éléments de nature aléatoire pouvant survenir durant un son voisé (i.e. bruit fricatif dans le phonème /v/) ou en l'absence de vibration des cordes vocales (i.e. bruit fricatif dans /s/, /t/, etc.)). Le but de ce travail est d'apporter une contribution à une modélisation précise de cette seconde composante et notamment des signaux de bruits modulés. Tout d'abord, une méthode de synthèse s'inspirant du bruit de grenaille, est introduite. Cette technique consiste à utiliser des processus ponctuels aléatoires qui définiront des instants d'occurrence d'événements spectraux (représentés par des Formes d'Ondes Formantiques ou FOF). Puis, s'appuyant sur un support théorique (représentation de Rice, théorie de la modulation aléatoire), un algorithme d'analyse/synthèse est proposé. Des tests de perception ont montré que cette méthode permet d'obtenir des signaux synthétiques jugés très naturels. De plus, cette approche apporte de nombreuses possibilités pour la modification de la qualité vocale (modifications temporelles, effort vocal, etc.).

Keywords: Speech decomposition; Aperiodic component of speech; Speech noises; Random formant wave forms; Analysis/synthesis; Rice representation; Speech modifications

1. Introduction

Early acoustic models of speech production made a clear separation between voiced excitation, due to regular vibration of the vocal cords, and unvoiced excitation. Obviously, this distinction does not hold for some types of speech sounds, such as voiced fricatives. This explains why mixtures of voiced and unvoiced excitation were introduced in formant speech synthesizers and also explains the success of Multi-pulse LPC (Atal and Remde, 1982). In this paper, the sound resulting from the quasi-periodic vibration of the vocal cords is referred as the "periodic component" of speech, and the sound resulting from aperiodic excitation (frication, aspiration, bursts) is referred as the "aperiodic component" of speech. The exact meaning of these terms is discussed in some details below.

In the context of concatenation-based synthesis (e.g. diphone synthesis), it is possible, using databases of natural speech, to capture voiced/unvoiced mixtures. However, methods for separately processing and modifying these components are desirable for high-quality systems, since modifications of the aperiodic component allow for voice quality transformation, or realistic prosodic modifications. The aim of the present work is to propose a method for analysis/synthesis of the aperiodic component of speech that allows new types of voice quality modification (in term of breathiness, roughness, hoarseness of a voice, time scale modification, frequency scale modification,...).

The analysis and synthesis of the speech aperiodic component has recently become a focus of interest for several reasons. On the one hand, this component seems responsible for a part of the perceived voice quality. Various studies aimed at associating the perceptual impression of a voice to simple parameters (see for example (Klatt and Klatt, 1990)). The ratio of the energy of the periodic component (or harmonic component) to the aperiodic (or noise) component is one of the useful parameters that can be used to describe hoarseness or breathiness in a voice (Hiraoka et al., 1984; Kojima et al., 1980; De Krom, 1993; Hillenbrand, 1987). On the other hand, some results (Laroche et al., 1993; Dutoit and Leich, 1993) indicate that a separate processing of the periodic and aperiodic components of speech signals may improve the quality of synthetic speech for time scale/pitch scale modifications.

In speech coding and speech synthesis, the aperiodic component is usually represented using a source/filter decomposition (Fant, 1960; Flanagan, 1972). Most of the models for speech synthesis use a Gaussian stationary excitation source and a slowly time-varying filter (Holmes, 1973; Liljencrants, 1968; Klatt, 1980; Rabiner,

1968), but only a few use amplitude modulation for the modelling of structural noises (noise in voiced segments) (Klatt, 1980; Rabiner, 1968). It is now acknowledged that it is important to take into account this time modulation of structural noises if one wants to obtain a good perceptual fusion between the periodic and the aperiodic component (Childers and Lee, 1991; Hermes, 1991; Chafe, 1990).

In analysis/synthesis systems, the stochastic part is usually modelled in the frequency domain exclusively (Rodet et al., 1987; Serra and Smith, 1990; Griffin and Lim, 1988; McAulay and Quatieri, 1992). But, because a white noise excitation source is not sufficiently precise in the time domain, (Laroche et al., 1993) proposes to time-modulate the noise by a time-domain energy envelope function. Though, this technique is specific to high pass filtered noises and cannot be successfully applied to wide band speech noises. Apart from these techniques, another method is introduced in (Marques and Abrantes, 1994) using Narrow-Band Basis Functions. These basis functions are obtained by amplitude modulating sinusoids by lowpass random processes chosen in a set of candidates (codebook). Nevertheless, there is no actual control on the signal time-domain evolution as the lowpass random process represents only the best candidate within a set of candidates. Furthermore, this approach does not provide perceptually relevant parameters (such as formants for example) that can be easily used for speech modification.

We, thus, decided to develop an algorithm in the framework of the source/filter decomposition but with an actual control on the time domain behavior of the signal. The filter is decomposed in several parallel formant filters excited by separate random sources. Within a formantic region, the passband noise signal is described as a random point process which defines the random times of arrival of the formant filter impulse responses (Formant Wave Form or FWF). Therefore, the aperiodic component is represented as a sum of elementary waveforms (the FWF) well localized in the spectro-temporal domain and related to meaningful acoustical features: the vocal tract resonances or formants. Such a representation allows for various types of usual speech modifications (time scaling, pitch, formant, plosive burst or fricative noise modification) but also for new types of speech modifications in term of breathiness, hoarseness or roughness of a voice.

The paper is organized as follows. In Section 2, the acoustic meaning of the aperiodic component and methods for periodic–aperiodic decomposition are discussed. In Section 3, the Random Wave Form synthesis concept is introduced. It is inspired by the "shot noise effect" and by the Formant Wave Form Synthesis. Section 4 describes the random modulation theory and Rice's representation which provide the theoretical framework of an original parameter estimation procedure. Section 5 provides an evaluation of the proposed algorithm by means of perceptual tests which include a comparison with two different techniques for the representation of the aperiodic component. The speech modification capabilities of the method are illustrated in Section 6 and some conclusions are finally suggested.

2. Periodic–aperiodic decomposition

2.1. What is the aperiodic component?

As mentioned above, various models used in speech synthesis or coding decompose the speech signal in two components: a periodic or quasi-periodic component which takes into account the quasi periodic segments of speech produced by regular vibrations of the vocal cords and an aperiodic component (or noise component). The aperiodic component corresponds to two main physical situations in speech production.

2.1.1. Additive noises

This source of aperiodicity represents different types of noises that are added to the periodic component.

1. *Transient noises.* They are, by nature, short and/or impulsives. This situation is encountered with stop consonants, where a rapid evolution of the articulators gives rise to an occlusion at a point inside the vocal tract followed by a sudden release of the air pressure. Mouth noises such as tongue clicks or lips noises are also put in this class.

2. *Quasi-stationary noises.* These noises correspond to the class of signals for which a turbulent flow appears at a constriction somewhere in the vocal tract (Stevens, 1971; Strevens, 1960), or at the glottis (Klingholz, 1987). This situation is encountered in whispered speech, in fricative or aspiration noise. Some quasi-stationary noises may also be found in bursts. The quasi-stationary speech noises have Gaussian amplitude probability densities (Richard et al., 1992) which is not the case for other types of speech signals whose amplitude probability densities are close to Gamma function (Davenport, 1952; Paez and Glisson, 1972).
3. *Modulated noises.* The turbulent flow generated at a constriction inside the vocal tract is modulated by the vocal cords vibration. This situation is encountered in voiced fricatives or in breathy vowels where noise is strongly time-modulated by the voiced source.

All these noises can be quoted "additive", because they correspond to another source of sound, that is added to the vocal cords sound source (when they are vibrating). Even if these noises may be strongly linked to the vocal cords source (consider for example the fricative noise modulated by the vocal cord air flow in the case of voiced fricatives), they correspond to a different source of sound, which is combined with purely voiced sound.

2.1.2. Structural noises

The noises produced by the random modulation of the amplitude, period and shape of the glottal waveform from period to period are known as structural noises (Klingholz, 1987). Different types of perturbations of the periodicity in source signals are:

1. *Jitter.* This is a random fluctuation of duration of fundamental periods.
2. *Shimmer.* This is a random fluctuation of amplitude for successive periods.
3. *Fundamental frequency variations.* Another source of aperiodicity, generally not random, is related to fundamental frequency (F_0) variations (i.e. glissando).

In the case of structural noise, periodic and aperiodic components are not combined by addition. This type of noise is linked to structural perturbation of the vocal cords vibration, and is not due to an additional sound source.

It should be noticed that these basic types of noise may combine in mixed types: transient and quasi-stationary noise in unvoiced stops, quasi-stationary and modulated noise in voiced fricatives, etc.

In normal voices, additive noise obviously represents the main contribution to the aperiodic component since the structural noises are usually low. However, to obtain the aperiodic component of natural speech signals, it is necessary to process the original signal by algorithms which inherently introduce some computational noise.

If this noise is negligible and speech synthesis databases contains low structural noises, the aperiodic component should contain only additive noises. In this case the decomposition into periodic and aperiodic components is theoretically correct. Then, aperiodic component modifications are meaningfull as they modify the speech source signal, which is closely related to voice quality.

2.2. Signal decomposition

Among the studies on the separation of the speech signal into a sum of two components (a deterministic (or periodic component) and an aperiodic component (or noise component)), some are devoted to the voice quality characterization. These studies do not explicitly separate the signal in two components but they rather measure a harmonic/noise ratio in order to describe different types of voice. In this framework, some studies are based on a frequency domain processing: Hiraoka et al. (1984) who use the relative intensity of the harmonics on the long term spectrum; Kojima et al. (1980) who reconstruct a long term spectrum from three periods of the original signal repeated indefinitely; or Klingholz (1987) who compares the long term spectrum with a harmonic spectrum deduced from the original signal. Some other studies work in the time-domain: Wendler et al. (1976), Yumoto et al. (1982) or Hillenbrand (1987) who average the signal periods on a long duration (> 1 s) and perform a comparison with the different successive cycles of the speech signal. Concurrently, another method

performs in the cepstral domain and leads to successful results (De Krom, 1993). Nevertheless, it appears difficult to adapt these methods to an explicit decomposition of the signal into two components. Actually, methods that seem to be well adapted for such a decomposition are those based on the sinusoidal model. Various models were proposed in this framework (Almeida and Silva, 1984; Griffin and Lim, 1988; McAulay and Quatieri, 1992; Rodet et al., 1987; Marques and Abrantes, 1994; Serra and Smith, 1990; Carl and Kolpatzik, 1991; George and Smith, 1992; Laroche et al., 1993; d'Alessandro et al., 1995a).

Among those methods, the method recently proposed in (d'Alessandro et al., 1995a) is preferred. This method is specifically designed for extracting the additive noise in speech, and thus provides a meaningful aperiodic component. Moreover, the ability of the algorithm for decomposing the additive noise and voiced excitation has been tested and demonstrated on both natural and synthetic signals containing a mixture of quasi-periodic excitation and noise excitation (d'Alessandro et al., 1995b).

The reader is reffered to (d'Alessandro et al., 1995a) for details on the algorithm, which is briefly summarized below. The method is based on spectral separation of periodic and aperiodic components, using an iterative signal reconstruction procedure (see Fig. 1).

The main steps are:

1. Separation of speech into an approximate excitation and filter components using Linear Predictive (LP) analysis. Periodic and aperiodic components are present in the excitation part of the speech production process. Thus, the aim of this first step is to obtain an approximation of the excitation signal. This signal is decomposed in short-duration frames.
2. A first identification of frequency regions of aperiodic and periodic components of excitation is performed,

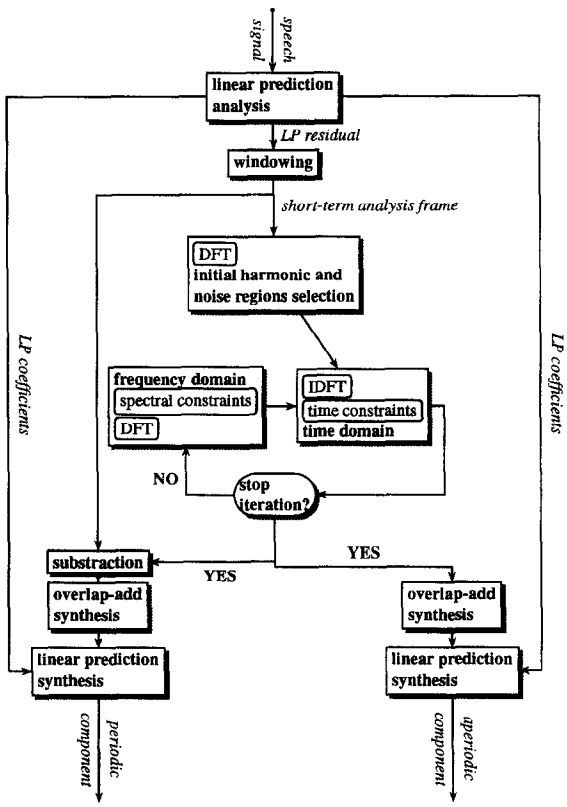

Fig. 1. Schematic diagram of the periodic/aperiodic decomposition algorithm.

for each frame. Using some knowledge on voicing and voice fundamental frequency, the periodic component is searched for in frequency (spectrum) and quefrency (cepstrum) domains. The frequency spectrum is divided in two regions: periodic regions (in the vicinity of harmonic frequencies) and aperiodic regions.

3. Using the first approximation obtained at the previous step, the two excitation components are built using an iterative reconstruction algorithm. The idea is to reconstruct complex aperiodic and periodic components from known samples in the regions obtained at the previous step, with an extrapolation algorithm, based on Fourier transform.
4. Finally, the periodic and aperiodic components of the excitation are then obtained in time domain by combining the reconstructed frames of data using an overlap-add procedure.
5. The periodic and aperiodic components are then passed through the time varying all-pole filter to obtain the components of the speech signal.

Fig. 2 illustrates the signal decomposition algorithm on a male voice.

An important question regarding the significance of the periodic and aperiodic components is whether these components represent some features of speech production or they merely are a convenient representation of speech signals. This question is discussed in detail in (d'Alessandro et al., 1995b) and it appears that:

- The periodic–aperiodic decomposition algorithm is able to separate additive random noise and periodic voicing for a wide range of fundamental frequency (F_0) variations. The dynamic range obtained (i.e. the average difference between the periodic component and the computational noise power spectra) is in all cases greater than 30 dB. The algorithm is able to separate continuous noise as well as pulsed noise.
- In the case of large jitter or shimmer values, both additive noise and structural noise are merged in the aperiodic component. Whilst it is still possible to achieve separation of a periodic and an aperiodic component, it seems difficult in this case to isolate the various production mechanisms of the aperiodic component. As such, the aperiodic component may be a useful parameter in the analysis of global voice quality, although it cannot be directly interpreted in terms of each underlying speech production parameters such as jitter of noise excitation amplitude. Finally, it has been noticed that the algorithm is almost not influenced by F_0 range or F_0 glides as this type of perturbation degrades only slightly the quality of the aperiodic component.

In conclusion, for speech synthesis applications where only high quality speech databases are used, the algorithm presented in (d'Alessandro et al., 1995a) provides a relevant aperiodic component that can be interpreted as additive noise.

3. Speech noise synthesis

In this section, a synthesis scheme for additive speech noises is introduced following a brief description of the underlying speech production model.

3.1. Speech signal model

According to the linear acoustic theory (Fant, 1960), speech production may be represented using a source/filter decomposition. Thus, the speech signal $s(t)$ may be expressed in time and frequency domains (denoted by capitals) as follows:

$$s(t) = e(t) * v(t) = (p(t) + ap(t)) * v(t), \tag{1}$$

$$S(\omega) = |S(\omega)|e^{j\theta_s(\omega)} \tag{2}$$

$$= (|P(\omega)|e^{j\theta_p(\omega)} + |AP(\omega)|e^{j\theta_{ap}(\omega)})|V(\omega)|e^{j\theta_v(\omega)}, \tag{3}$$

where $s(t)$ is the speech signal, $v(t)$ is the impulse response of the vocal tract system, $e(t)$ is the excitation signal, $p(t)$ is the quasi-periodic part of the excitation, $ap(t)$ is the random part of the excitation, and θ symbolizes the phase of the corresponding component.

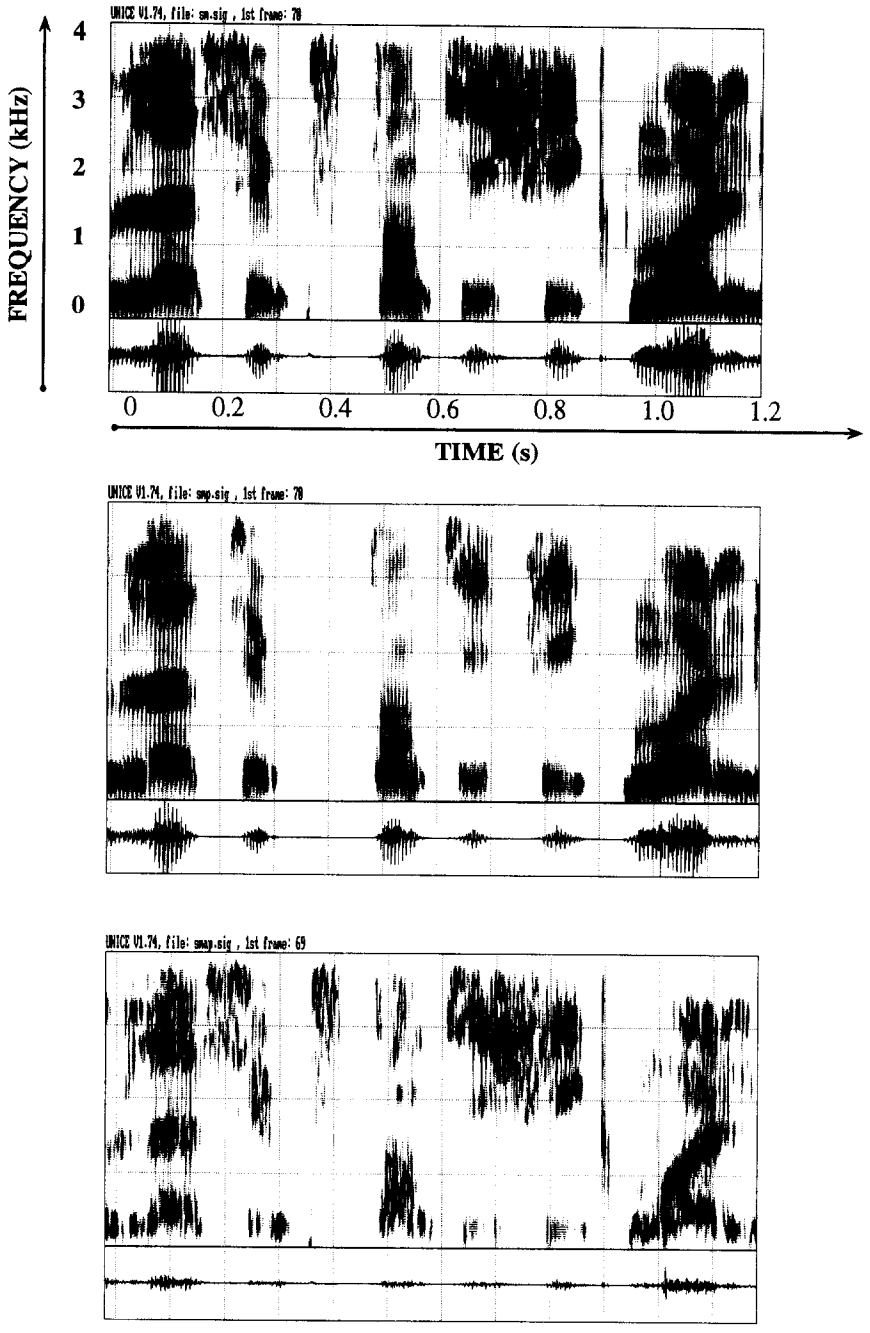

Fig. 2. Wide-band spectogram (male voice). "la substantifique moelle ...". Top: original speech. Middle: periodic component. Bottom: aperiodic component. Audiofiles available (http://www.elsevier.nl/locate/specom).

In such a model, a speech noise $b(t)$ is described as an excitation source $ap(t)$ filtered by a time-varying filter v with impulse response $v(t)$:

$$b(t) = ap(t) * v(t). \tag{4}$$

The excitation source $ap(t)$ represents the acoustic noise due to the turbulent flow created at constriction points inside the vocal tract or at the glottis. The time-varying filter v represents the action of the vocal tract on this excitation source. In the frequency domain, the power spectrum $S_{bb}(\omega)$ of a speech noise $b(t)$ can be factorized into two components:

$$S_{bb}(\omega) = S_{apap}(\omega)|V(\omega)|^2, \tag{5}$$

where $S_{apap}(\omega)$ is the power spectrum of the noise source and $V(\omega)$ is the frequency response of filter v.

3.2. Frequency-domain decomposition of the vocal tract response

One of the most important features of speech spectra is that they possess several perceptually relevant maxima (or formants), related to the different resonances of the vocal tract for a particular geometrical configuration. Therefore formant-based methods are commonly used for the synthesis of unvoiced speech. The parallel decomposition is generally preferred for unvoiced speech synthesis as a precise control on each formant amplitude is possible and as the main drawback of parallel decomposition (interferences between neighboring branches) does not stand in the case of a random excitation because of the irrelevance of the phase spectrum (Flanagan, 1972; Holmes, 1983; Klatt, 1980).

Thus, the vocal tract filter may be approximated by a set of second order resonators v_i. In time domain, this spectral decomposition can be written:

$$v(t) = \sum_{i=1}^{M} v_i(t - t_i), \tag{6}$$

where $v_i(t)$ represents the impulse response at time t, of the ith parallel section excited at time t_i. Using Eqs. (4) and (6), it is deduced that

$$b(t) = ap(t) * \sum_{i=1}^{M} v_i(t - t_i). \tag{7}$$

$v_i(t)$ represents a resonance of the vocal tract in the time-domain and defines a Formant Wave Form (FWF):

$$v_i(t) = A_i e^{-\alpha_i t} \sin(\omega_i t + \phi_i). \tag{8}$$

In this work, the Formant Wave Forms chosen are those introduced by Rodet (1980):

$$v(t) = \Lambda(t) e^{-\alpha t} \sin(2\pi f_c t + \phi), \tag{9}$$

with

$$\Lambda(t) = \begin{cases} 0 & \text{if } t \le 0, \\ \frac{1}{2}A(1 - \cos(\beta t)) & \text{if } 0 < t \le \pi/\beta, \\ A & \text{if } t > \pi/\beta. \end{cases} \tag{10}$$

A FWF is thus characterized by the following parameters:

- π/β, the excitation time which is defined as the duration between the onset of an FWF to the maximum of its envelope,
- f_c, the formant center frequency,
- α/π, which represents both the -3 dB power spectrum bandwidth in the frequency domain and the rate of waveform damping in the time-domain,
- A, the formant amplitude,
- ϕ, the initial phase.

Compared to impulse responses of second order resonators, the FWF have an additional excitation time parameter in the envelope $A(t)$. Practically, this parameter is particularly useful as it can be used along with the formant bandwidth to control the time domain envelope of an FWF. Depalle (1991) shows that it can be linked to the bandwidth of the resonator at -12 dB, and is often called "width of the skirt". It is shown in (d'Alessandro, 1989, pp. 160–163) that this parameter can also be interpreted as the width of the frequency windows where the effect of the FWF is significant.

3.3. Random formant wave form synthesis

The Formant Wave-Form synthesis method has initially been introduced for voiced speech synthesis (Rodet, 1980). In the original implementation, the FWF are periodically generated in order to obtain periodic sounds as voiced speech, singing voices or various musical instruments. In a previous work (d'Alessandro, 1990), it is experimentally shown that unvoiced speech can be synthesized using FWF generated at random points in time. Consequently, an original synthesis process in the framework of shot noise is proposed and is briefly described below.

The *shot noise effect* is an important example of random point process in physics and is due to fluctuations in the intensity of the stream of electrons flowing from the cathode to the anode of a vacuum tube. Let t_i be the time of arrival of the ith electron at the anode and let $h(\tau)$ be the effect on the current of one electron, the total current at time t may be represented by

$$s_t = \sum_{i=1}^{N} h(t - t_i). \tag{11}$$

The process s_t is commonly called shot noise and has been thoroughly studied by several authors (Rice, 1944–1945; Davenport and Root, 1958; Snyder, 1975). They all agree that the random point process (t_i for $i \in [1, N]$) may be reasonably modelled as a Poisson process. Thus, shot noise is represented as a filtered Poisson point process.

In previous work, it is shown that Gaussian noise can be synthesized using the framework of shot noise (Richard et al., 1992). As a matter of fact, quasi-stationary speech noises such as unvoiced fricatives have Gaussian amplitude probabilities and it is important to keep this property to model properly this class of noises. Concurrently, it has been noted that synthetic noises of this class sound more smooth and natural when they have Gaussian amplitude probabilities (Depalle, 1991, p. 92).

The idea is to generate a shot noise in each formant region by replacing the filter $h(t)$ by Formant Wave Forms v_i, and the random excitation $ap(t)$ in Eq. (7) by a random point process:

$$ap(t) = \sum_{j} \delta(t_{ij}), \tag{12}$$

where t_{ij} represents the jth random point (or impulse) of the ith formant point process.

The aperiodic component is then synthesized using the following equation:

$$b(t) = \sum_{i=1}^{M} \sum_{j} v_i(t, t_{ij}), \tag{13}$$

where M is the number of formants, where t_{ij} represents the jth random point (or impulse) of the ith formant point process, and where $v_i(t, t_{ij})$ is the FWF generated at time t_{ij}.

The synthetic noise $b(t)$ approaches a Gaussian process if the densities (density should be understood as the average number of impulses per unit time) of the point processes are large compared to the effective duration of the impulse responses: in other words, when each sample of unvoiced speech is the sum of a large number of impulse responses. A demonstration of this general result can be found in (Papoulis, 1986, pp. 629–635). If this

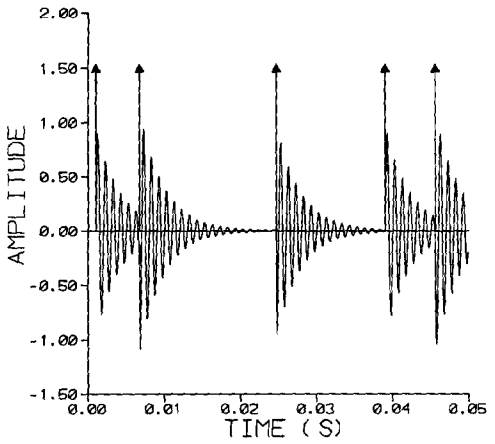

Fig. 3. Low density synthesis scheme using only one formant. The arrival of a new FWF does not cut the preceding one but are added to give the resulting synthetic signal. Obviously, to obtain Gaussian synthetic signals, the density of points must be much higher (Richard et al., 1993).

condition is reached, the output signal is Gaussian, and its power spectral density is imposed by the amplitude spectra of the FWF. In this case, it is therefore equivalent to filtered Gaussian white noise. Fig. 3 gives an example of low density random FWF (RFWF) synthesis, using only one formant.

By extension, the FWF parameters can be time varying and the point process can range from quasi-periodic to purely random which allows to model all classes of noises. In practice, the parameters are fixed for the duration of a single FWF, but successives FWF can have different parameters.

Given the synthesis parameters, the Random Formant Wave Form synthesis is performed in three steps:

1. A set of random points is defined.
2. FWF are generated according to the (deterministic) acoustic parameters (formant center frequency, amplitude, bandwidth, initial phase and the excitation duration (π/β), and according to the random points (which define the times of arrival or instants of generation).
3. The RFWF signals computed in the different branches are summed together.

Several procedures have been proposed for an automatic estimation of the synthesis parameters (Richard et al., 1993; d'Alessandro, 1990; Liénard, 1987). A new automatic estimation technique working in the time domain is proposed below.

4. Analysis / synthesis of the aperiodic component

In this section, an analysis/synthesis method for the aperiodic component is proposed. The theoretical framework of this technique is provided by the random modulation theory and Rice's representation and is briefly described below. This theory has many interesting properties and, in particular, it allows to define an unambiguous time-domain envelope and instantaneous phase from which the parameters of RFWF synthesis are deduced.

4.1. Random modulation theory and Rice's representation

In this section, we consider that $x(t)$ is a bandpass stochastic signal. A basic topic in random modulation is the representation of such processes as an amplitude modulated signal. This representation describes any

bandpass stochastic signal $x(t)$ as a random (real) envelope $r(t)$ modulating an oscillating term:

$$x(t) = r(t)\cos[\psi(t)] = r(t)\cos[\omega_0 t + \phi(t)]. \tag{14}$$

Though, because an infinity of pairs $[r(t), \psi(t))]$ can be associated to a given signal, it is decided to use the couple deduced from the analytical signal $z(t)$:

$$z(t) = x(t) + j\hat{x}(t) = r(t)\,e^{j(\omega_0 t + \phi(t))} = v(t)\,e^{j\omega_0 t}, \tag{15}$$

where $\hat{x}(t)$ denotes the Hilbert Transform of $x(t)$. This choice, for the definition of the envelope and the instantaneous phase, corresponds to the *Rice Representation* (Rice, 1944–1945; Papoulis, 1983, 1986) and unambiguously defines the envelope and the instantaneous phase of $x(t)$ (Picinbono and Martin, 1983; Picinbono, 1989).

It is possible to show that the envelope $r(t)$ is independent of the carrier frequency ω_0. Furthermore, Rice's representation is valid for any probabilistic model: $x(t)$ may be stationary, cyclostationary or non-stationary.

In the case of Wide Sense Stationary signals, Rice's representation corresponds to an optimal solution in the sense of the minimization of the average rate of change of the envelope $r(t)$. It is possible to show (Mandel, 1967; Papoulis, 1983) that Rice's representation is equivalent to the minimization of the following integral which defines a measure of this rate:

$$I = E\left[|v'(t)|^2\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega^2 S_{vv}(\omega)\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} (\omega - \omega_0)^2 S_{zz}(\omega)\,d\omega, \tag{16}$$

where $v(t) = r(t)e^{j\phi(t)}$ is the complex envelope, and S_{vv} represents the power spectrum of signal v.

Some properties may then be deduced:

1. *On the carrier frequency* ω_0

- the optimal carrier frequency $\overline{\omega_0}$, in the sense of the minimization of I, is the center of gravity of the power spectrum of $x(t)$:

$$\overline{\omega_0} = \frac{\int_0^{\infty} \omega S_{xx}(\omega)\,d\omega}{\int_0^{\infty} S_{xx}(\omega)\,d\omega}; \tag{17}$$

- it is also the weighted average of the instantaneous frequency $\omega_{\text{inst}}(t)$ of $x(t)$:

$$\overline{\omega_0} = E\left[r^2(t)\,\omega_{\text{inst}}(t)\right]/E\left[r^2(t)\right], \tag{18}$$

where the instantaneous frequency is given by

$$\omega_{\text{inst}}(t) = 2\pi f_{\text{inst}}(t) = \frac{x(t)\hat{x}'(t) - x'(t)\hat{x}(t)}{r^2(t)}, \tag{19}$$

where x' denotes the time derivative of x.

2. *On the real envelope* $r(t)$

- the envelope is maximally smooth (the envelope fluctuations are as slow as possible).

This last property makes it possible to study the temporal characteristics of the envelope local maxima and to built from their positions a point process as shown in Fig. 4. The idea is to use this point process as the virtual excitation source of the RFWF synthesis. The nature of this point process is obviously dependent on the studied signal as modulated noises would lead to almost periodic point processes and quasi-stationary noises to purely random processes. Random point processes have been thoroughly studied in the past as they can describe many

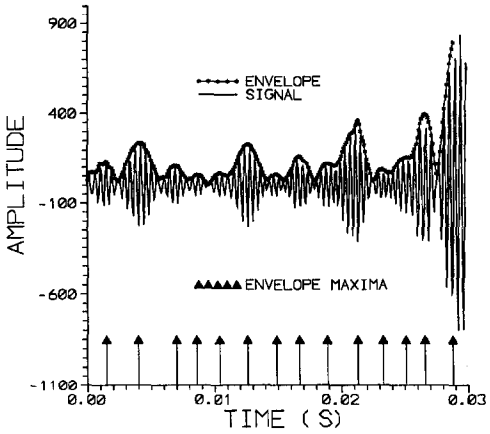

Fig. 4. The point process impulses are defined from the (time-domain) envelope maxima location (from (Richard et al., 1993)).

physical phenomena (Shot noise, lightning discharges, Astronomy, seismic events, etc.) (see (Snyder, 1975; Leadbetter, 1972)). Rice (1944–1945), in his pioneering work, studied the distribution of the maxima of the envelope of shot noise and showed that analytical expressions may be found for extremely simple case (symmetric and ideal band pass filter under the assumption of Wide Sense Stationary processes). Despite its interest, the detailed characterization of the various point processes obtained from speech noises analysis is a delicate problem and is beyond the scope of this paper.

4.2. *From amplitude modulation to the FWF representation*

For natural speech, the Amplitude Modulation representation needs to decompose the signal in a rather high number of bands to obtain high quality synthetic signals (Flanagan, 1980). This is not the case for the aperiodic component of speech where only a few bands are necessary (approximately one per formant region).

Let $x_b(t)$ be a bandpass stochastic signal (i.e. a signal obtained by filtering the original aperiodic signal in a formant region). This signal may be written in the form of Eq. (14):

$$x_b(t) = r(t)\cos\left[\omega_0 t + \phi(t)\right]. \tag{20}$$

To obtain the RFWF representation, it is first considered that the temporal envelope $r(t)$ may be decomposed into a sum of short duration envelopes referenced at the time instant t_i:

$$r(t) = \sum_i r_i(t - t_i). \tag{21}$$

Eq. (20) then becomes

$$x_b(t) = \sum_i r_i(t - t_i)\cos\left[2\pi f_0 t + \phi(t)\right]. \tag{22}$$

Secondly, it is assumed that the oscillating term has a constant initial phase: $\phi(t) = \psi$ for the duration of the envelope $r_i(t)$. This corresponds, in a first approximation, to a hypothesis of signal stationarity for the duration of an envelope $r_i(t)$. In other words, this assumption is equivalent to consider a constant filter between two successive maxima. This is a less stringent hypothesis than the usual quasi-stationary assumption of speech

signals since an FWF duration usually is shorter than a conventional analysis window length (which is typically 10 ms).

Then, Eq. (22) gives

$$x_b(t) = \sum_i r_i(t - t_i)\cos(2\pi f_0 t + \psi_i), \tag{23}$$

$$x_b(t) = \sum_i r_i(t - t_i)\cos(2\pi f_0(t - t_i) + \phi_i), \tag{24}$$

with

$$\phi_i = \psi_i + 2\pi f_0 t_i. \tag{25}$$

The FWF parameters are then estimated in two steps: (1) the estimation of the envelope parameters; (2) the estimation of the center frequency and initial phase.

4.3. RFWF envelope parameters estimation

By using Eqs. (9), (10) and (24), one obtains the envelope $r_i(t - t_i)$ as a function of the FWF envelope parameters:

$$r_i(t - t_i) = \begin{cases} 0 & \text{if } (t - t_i) \le 0, \\ \frac{1}{2}A_i\left(1 - \cos(\beta(t - t_i))\right)e^{-\alpha(t - t_i)} & \text{if } 0 < (t - t_i) \le \pi/\beta, \\ A_i e^{-\alpha(t - t_i)} & \text{if } (t - t_i) > \pi/\beta, \end{cases} \tag{26}$$

where t_i defines the instant of generation of the ith Formant Wave Form of the synthetic signal and also the ith envelope minimum location of the analyzed bandpass signal. Thus, by fitting the FWF envelope, with the analyzed signal envelope between two successive minima $[t_i = t_{\min 1},\ t_{i+1} = t_{\min 2}]$ (Eq. (26)), one obtains the following FWF parameters:

- $T_{ex} = \pi/\beta$: the excitation time defined as the duration between an envelope minimum $t_i = t_{\min 1}$ to the next envelope maximum $t = t_{\max}$.
- $\alpha = \pi l_b$ which represents both the -3 dB power spectrum bandwidth in the frequency domain and the rate of waveform damping in the time-domain. This parameter is obtained from Eq. (26) (3rd line) taken at time $t = t_{\max}$ (at the envelope maximum) and at time $t = t_{\min 2}$ (at the next envelope minimum).
- A_i which represents the amplitude of the waveform.

4.4. FWF center frequency estimation

The FWF center frequency is deduced from the instantaneous frequency of the signal $x_b(t)$. Several techniques may be used to obtain an estimation of the instantaneous frequency (for example (Papoulis, 1983; Tsopanoglou et al., 1993; Berthomier, 1983; Grandsten and Beet, 1993)). The solution given by the random modulation theory (see Section 4.1) is preferred, as an estimation of the signal envelope in the analysis procedure is already available.

The FWF center frequency f_c is obtained as the optimal frequency $f_c = \overline{\omega_0}/2\pi$ in Eq. (14). It is the weighted average of the instantaneous frequency (see Eq. (18)):

$$f_c = \frac{\sum_{t\in[t_{\min 1}, t_{fwf}]} r^2(t - t_i) f_{inst}(t - t_i)}{\sum_{t\in[t_{\min 1}, t_{fwf}]} r^2(t - t_i)}, \tag{27}$$

where $[t_{\min 1},\ t_{fwf}]$ represents the FWF duration interval and where f_{inst} represents the instantaneous frequency (estimated by Eq. (19)).

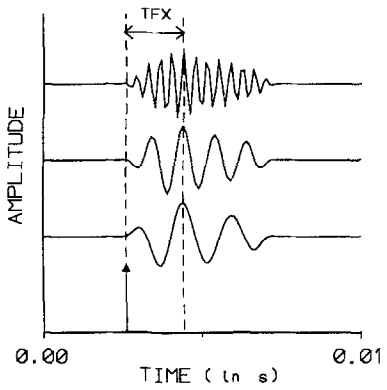

Fig. 5. The initial phase is set as a function of formant center frequency. In this ideal example, the excitation time parameter (T_{ex}) which defines the time between the beginning of the FWF to the maximum of its envelope is kept constant.

The FWF initial phase is finally determined as a function of the FWF center frequency in order to give a maximum at the exact place defined by the envelope maximum. This is essential if one wants to keep the modulated structure of some structural noises (see Fig. 5)

Once the parameters of an FWF are estimated, the FWF contribution outside the analysis window (i.e. for $t > t_{min2}$) is subtracted from the temporal envelope of the original signal $x_b(t)$. The analysis procedure may then be iterated until the end of the signal is reached.

It must be noted that the frequency bands are predefined once for all and do not change with time. In other words, no formant detection is performed to define those bands. Nevertheless, within each predefined band, an explicit formant extraction is done using the algorithm described above. If no formant falls in a band, the amplitude of the corresponding RFWF will be very low or null. If two formants fall in the same band, they will be modelled as a single but wider formant which has an almost negligible effect for speech noises. Although there is no overall formant tracking, it is possible, by combining the parameters obtained in each band, to visualize the trajectories of each formant (see below Fig. 10).

The general scheme of this analysis/synthesis algorithm is given in Fig. 6. Additional details on the algorithm may be found in (Richard, 1994).

5. Evaluation of the results

5.1. Probability density of quasi-stationary speech noises

Because quasi-stationary speech noises such as unvoiced fricatives (i.e. /f/, /s/, etc.) have Gaussian probability densities, the RFWF method is first tested on his ability to generate noise with this property. The Gaussian property is important to synthesize more natural frication and aspiration noises (Depalle, 1991).

The algorithm described above is run on sustained realization of the three unvoiced French fricatives (/f/, /s/ and /ʃ/). Thus, for this experiment, the speech decomposition algorithm is not used since unvoiced fricatives may be characterized as aperiodic signals. Using a χ^2 test (Foucart, 1991, pp. 123–148), Gaussian probability densities are obtained for the three unvoiced fricatives.

The χ^2 test allows to say if the difference between an empirical probability density and a theoretical Gaussian curve is significant or not. For the three (French) unvoiced fricatives /f/, /s/ and /ʃ/, the statement "the difference between their empirical probability densities and the closest theoretical Gaussian curve is insignificant" was true with an error risk inferior to 0.5%.

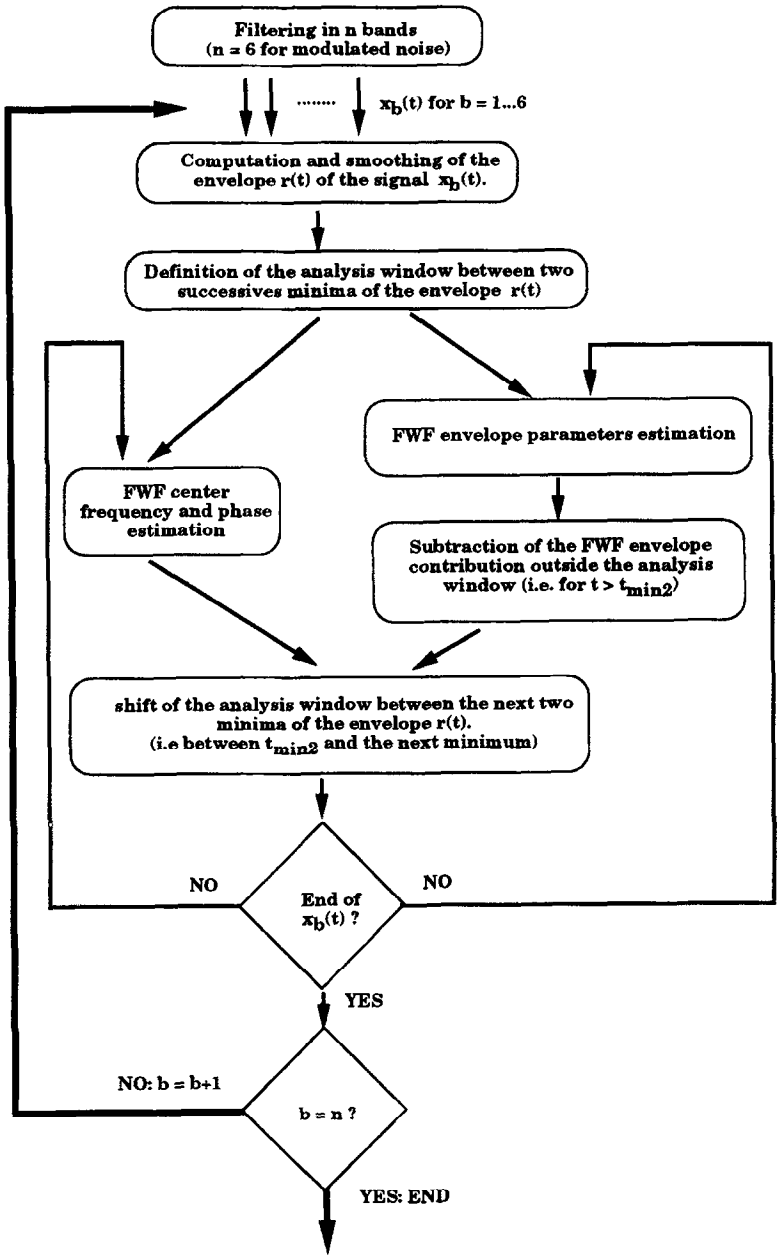

Fig. 6. General block diagram of the FWF analysis/synthesis method.

Fig. 7 shows the logarithm of the probability density for the unvoiced fricative (/ʃ/) and for its resynthesized version.

5.2. Perceptual evaluation

To evaluate the RFWF method, it is decided to compare it with two other algorithms: the traditional Linear Predictive (LP) analysis/synthesis model (or LPC (Atal and Schoeder, 1970)) and the more recent modulated

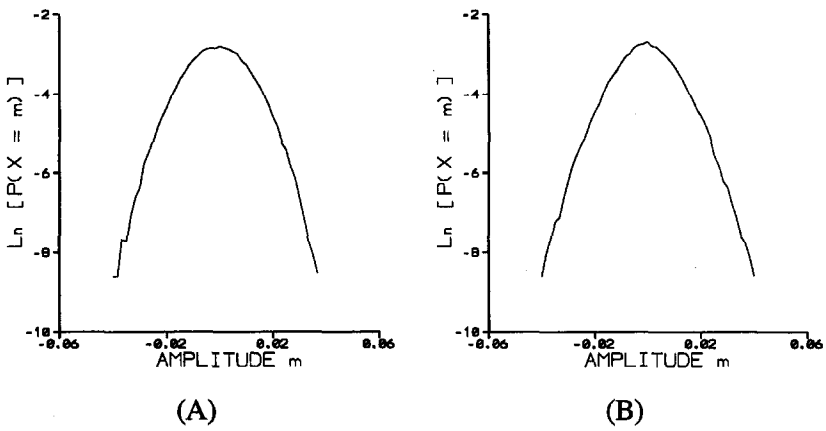

Fig. 7. (A) Logarithm of the empirical probality density for the unvoiced fricative /ʃ/. (B) Logarithm of the empirical probability density for the unvoiced fricative /ʃ/ resynthesized by the FWF model. The amplitude m is given relatively to the maximal amplitude value authorized by analog/digital converter (from (Richard et al., 1992)).

LP analysis/synthesis representation (referred herein as modulated LPC, mLPC) introduced in the Harmonic + Noise Model (see (Laroche et al., 1993)). Even if the traditional LPC model does not represent the intrinsic modulation of speech noises, it is certainly the most widely used and known model for modelling the aperiodic component of speech. The other algorithm (mLPC) models the aperiodic component of speech as a time-modulated signal and is chosen for comparison because of its simplicity. In this model, a low-pass filtered envelope of the signal is used for modulating a white noise filtered by the LPC filter.

Fig. 8 gives the results of the three algorithms (FWF, LPC and mLPC) in the time domain.

It appears that the noise temporal structure is well represented by the RFWF and the mLPC methods, whereas it is not using traditional LPC. However, it appears that a more accurate time-domain description is obtained with the RFWF method. As a matter of fact, LPC cannot trace the modulated structure that is present in the aperiodic component.

To make an objective evaluation of the perceptual performance of the RFWF algorithm, it is decided to run formal perceptual evaluation tests. Because the mLPC algorithm has been initially introduced for high-pass speech noises and cannot be successfully applied to wide band speech noises, two different tests are proposed. The first test compares the three algorithms on wide band aperiodic components, whereas the second uses only high pass filtered version of the original aperiodic component (typically only frequencies above 1kHz are kept).

This allows us to obtain:

- an evaluation of all algorithms on realistic aperiodic components, i.e. that possess also low-pass frequencies (Test 1);
- a more fair evaluation with the mLPC method which is not particularly fitted for representing wide band modulated noises (Test 2).

The experimental paradigm chosen for perceptual testing is the classical Degrading Category Rating test (DCR). A description of this well known test can be found in (CCITT, 1992).

Sixteen subjects (including one of the authors), most of whom are familiar to these types of tests, are asked to give an appreciation of the degradation of synthetic signal (second of a pair) compared with natural signal (first of the pair). The subjects are all laboratory members, and some of them have already participated in perception experiments. All subjects have normal hearing.

The answers have to be chosen within the five following categories:

1. the degradation is very annoying,
2. the degradation is annoying,

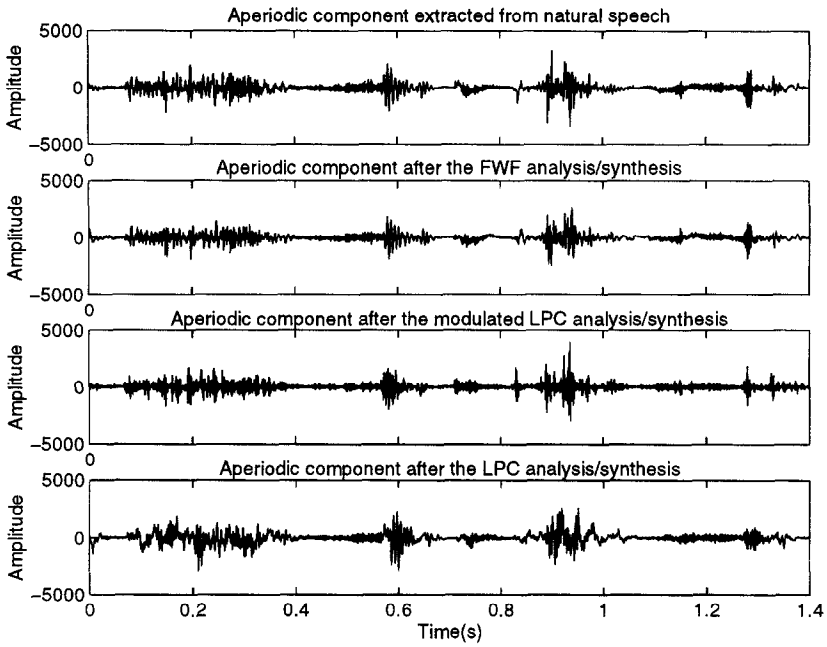

Fig. 8. Time-domain signal of: (top) the original aperiodic component of speech; (middle-up) the aperiodic component obtained by the FWF analysis/synthesis model from the original aperiodic component; (middle-down) the aperiodic component obtained by the LPC analysis/synthesis model from the original aperiodic component; (bottom) the aperiodic component obtained by the modulated LPC analysis/synthesis model from the original aperiodic component.

3. the degradation is slightly annoying,
4. the degradation is audible but not annoying,
5. the two signals are equal.

The test corpus consists of 6 sentences (randomly chosen in a speech corpus of read text) of at least 2 s duration (3 males/3 females). It must be emphasized that these sentences are randomly chosen with no special care about accent, age, speaking rate or specific voice characteristics.

The periodic and aperiodic parts are separated and, for each test, four different pairs are built for each sentence:

S1. The "reference pair". The two members of the pair are identical: the natural speech signal (the sum of the periodic and aperiodic components).

S2. The pairs associating the natural speech and the reconstructed speech obtained by the addition of the periodic component and the stochastic component modelled using LPC.

S3. The pairs associating the natural speech and the reconstructed speech obtained by the addition of the periodic component and the stochastic component modelled using mLPC.

S4. The pairs associating the natural speech and the reconstructed speech obtained by the addition of the periodic component and the stochastic component modelled using RFWF.

According to the DCR procedure, it is not fair to built a pair associating two synthetic signals since it would have implied that the first method outclasses the second one.

To test the robustness of the algorithms, it is also decided to compare the three methods on signals obtained by modification of the aperiodic component. These modifications are simple changes in the amplitude of the aperiodic component, scaled by a factor 2 or 3.

The aim of this test is to measure the degree of fusion of the aperiodic and the periodic components and to

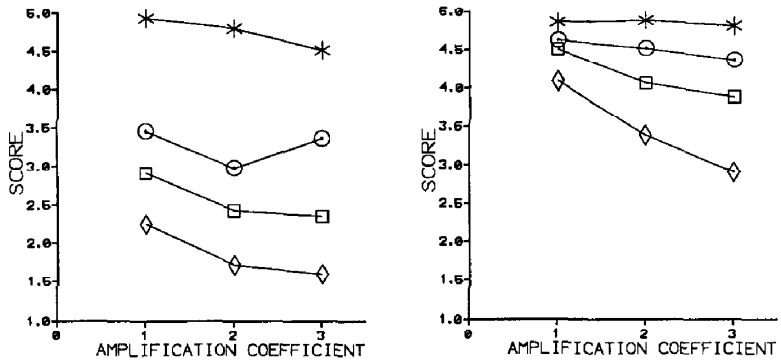

Fig. 9. Perception test results (Test 1 (left) and Test 2 (right)). *X*-axis: amplification coefficient of the aperiodic component. *Y*-axis: average DCR score. A score of 5 corresponds to the answer "the signals within a pair are equal", and a score of 1 corresponds to the answer "the degradation in the second signal is very annoying"). Stars are for pairs of identical signals. Circles are for pairs where the second signal is reconstructed using the random FWF method. Squares are for pairs where the second signal is reconstructed using LPC and diamonds are for pairs where the second signal is reconstructed using the modulated LPC.

test the robustness of this method when the aperiodic component is modified. Consequently, 72 different pairs are obtained for each test (6 sentences × 4 algorithms × 3 scales for the aperiodic component). Two different tests are run, for full-band aperiodic components and high-pass filtered aperiodic components. Therefore 144 stimuli are presented to the subjects. [3]

The tests take place in a sound-insulated booth. Stimuli are presented binauraly through Beyer DT48 headphones at a level of 80 dB SPL. Stimuli are played in a random order, and it is possible for the subjects to listen to a stimulus pair as many time as they want, before reporting an answer. An experimental session lasts half an hour.

The results of the DCR tests are given in Fig. 9. The results suggest that:

1. When wide band aperiodic signals are used, the FWF clearly outclasses the other methods. This also means that taking into account the modulated structure of speech noises is important to obtain a good fusion between the two speech components. We think that these results are linked to the better time and frequency accuracy of our method: formants are well represented, and the time domain control gives a better perceptual fusion between the periodic and aperiodic components. It also seems that the formants in noise are represented with a better accuracy using the FWF method, compared to both LPC and mLPC algorithms which shows that predefining the frequency bands without estimating formant trajectories is appropriate for speech noises.
2. When only high pass speech noises are used, all methods obtain results that correspond to a not annoying degradation when audible (for no amplification coefficient). This tends to show that there is no need to incorporate a precise time-domain control for such signals (high pass noises). However, when the aperiodic component is amplified, the LPC performances degrades significantly compared to the FWF algorithm.
3. Amplification of the aperiodic component introduces a perceptual degradation, even for natural components (S1). However, this is not true for the high-pass filtered conditions.
4. The mLPC method shows poor performances for full-band signals. For implementation of the mLPC method, we followed the instruction directly provided by one of the authors of (Laroche et al., 1993). Even for high-pass filtered signals, LPC seems slighly better than mLPC. This surprising result can probably be explained by the fact that mLPC is very sensitive to errors in speech decomposition. As a matter of fact, the

[3] Audiofiles available (http://www.elsevier.nl/locate/specom).

mLPC algorithm is particularly affected by the presence of structural noise in the aperiodic component after decomposition.

5. All the synthetic stimuli S2, S3, S4 introduce audible artefacts when compared to the original signal. But in the most favorable case, these differences are judged not annoying.

This perceptual experiment demonstrated that the RFWF method performs better than other known methods for representing the speech aperiodic component. However, synthetic and natural signals are distinguishable, even if the difference is not annoying. In case of simple modifications of the aperiodic component, the RFWF method also shows better performances.

5.3. Complexity of the algorithm

Although the synthesized noise quality and naturalness is better with our method than with other models, the complexity (both in terms of computation and data rate) is higher than classical LPC. However, it does not seem excessive for practical synthesis applications.

The experiments show that for each synthesis sample it is necessary to add the contribution of about 10–20 FWF, where each FWF is the product of a sinusoidal component and an envelope component.

The analysis algorithm is more intricate. It involves bandpass filtering and FWF parameters estimation. But this analysis stage is usually done off-line. It does not need to be real time for most speech synthesis applications.

Typically (for signals sampled at 8kHz), the number of FWF needed for synthesis is of the order of one thousand per second which leads to a rate of about five thousand parameters per second. Furthermore, this data rate can be easily lowered by suppressing the low energy FWF as more than 50% of FWF are nearly not audible because of their very low amplitudes. If one considers that the parameters do not need to be coded using large words (an average of 7–8 bits per parameter seems sufficient), the method achieves a data reduction compared to the signal data rate (35kbytes/s versus 128 kbytes/s for 16 bits, 8 kHz signals).

However, the aim of this study is to design a technique which is able to represent noise with accuracy, and not to perform a parameter rate reduction. RFWF synthesis, with or without modifications, is easily performed in real time on a modern workstation.

6. Application to speech modifications

One of the advantages of the proposed method is that many types of modifications of the aperiodic component are almost trivial, and result in high-quality synthetic signals.

Although it is possible to separate a periodic and an aperiodic component, many voice quality modifications affect both components. However, only modifications of the aperiodic components are discussed below.

At the output of the analysis stage, the aperiodic component is represented by a set of elementary waveforms described by relevant acoustic parameters. Formant center frequency, bandwidth and amplitude, and excitation time have a direct impact on the power spectrum and on the time-domain envelope and appear particularly useful to carry speech modifications. This is also true for excitation time, because this parameter can be related to the spectrum bandwidth at -12 dB of the peak.

As it can be seen in Fig. 10, the aperiodic component is decomposed by the analysis-synthesis algorithm into a set of short-term narrow-band functions (represented as circles). A parametric description of these functions is available in a file containing FWF parameters, which represents a decomposition of the signal in time and frequency. Speech modifications are achieved by simple manipulations of these FWF parameters. Synthesis is performed by the same synthesis algorithm that is used in the analysis-synthesis system.

Very simple modifications of the FWF parameters give rise to high-quality modifications of relevant speech attributes, such as speech rate, vocal effort and voice quality.

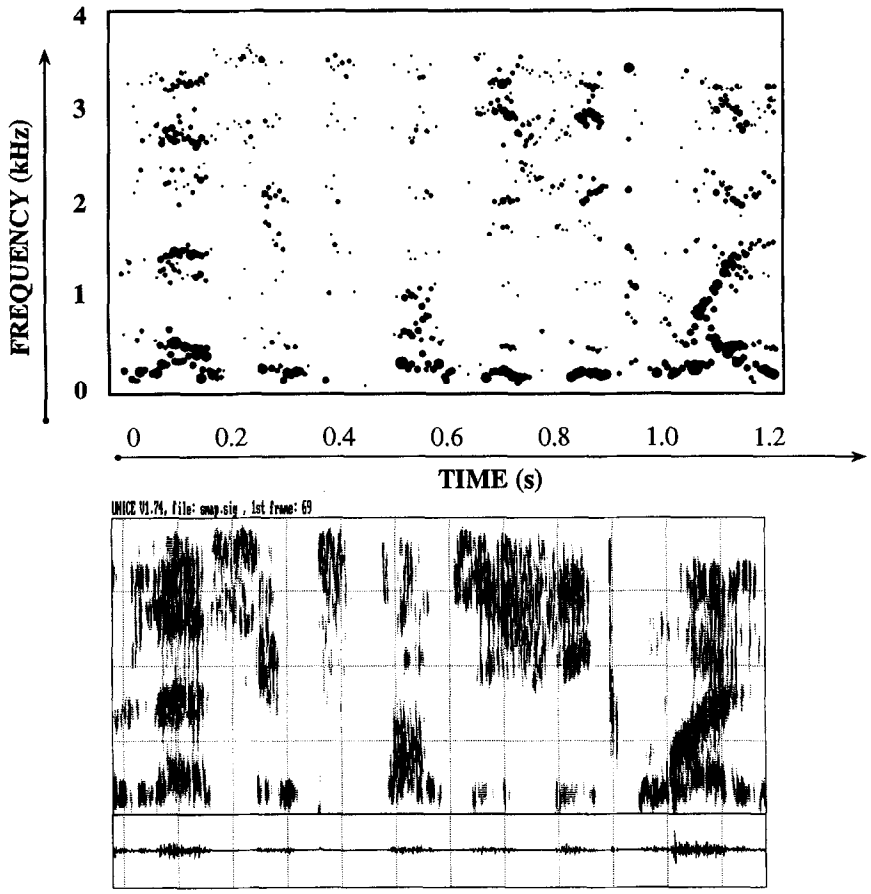

Fig. 10. Top: representation of the RFWF in the time-frequency domain. Each FWF is schematically plotted as a circle, whose radius corresponds to the FWF logarithmic energy. The location of each circle is determined by the FWF center frequency and time of generation. Bottom: corresponding wide-band spectrogram (same signal as Fig. 2).

6.1. Time scaling

FWF are referenced in time by the instant of generation. By simply modifying this parameter, high-quality time scaling of small coefficients (0.5–2) can be performed. The results obtained are fairly good either for compression and dilatation. Speeding the speech rate of a factor of 2, for example, is achieved simply by dividing all the instants of generation by 2 before synthesis (see Fig. 11).

However, for large dilatation coefficients (greater than a factor of 2) more sophisticated procedures are needed: each FWF must be duplicated in time with a certain amount of randomness to avoid tonal quality in the dilated segment. This type of time scaling results in a global dilatation or compression without affecting the (possible) underlying periodicity of noise modulation.

This modification procedure is rather similar to other modification procedure, for instance the PSOLA (Pitch Synchronous OverLap Add) algorithm (Moulines and Charpentier, 1990), but it does not require pitch markers. It is similar to PSOLA because modifications are carried out merely by changing times of occurence of short duration signals extracted from the natural signal. But contrary to PSOLA, the short duration signals are described by a parametric representation, and are also narrow band signals.

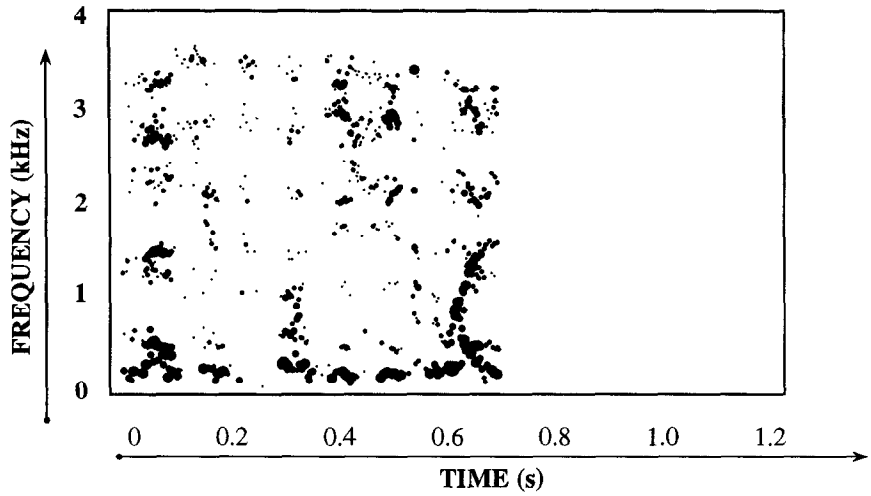

Fig. 11. The speech signal displayed in Fig. 10 is compressed by a factor 2. Audiofiles available (http://www.elsevier.nl/locate/specom).

6.2. Voice quality modification

Because the proposed representation is based on a formant description, various straightforward spectral modifications can be performed. By altering the center frequencies, it is possible to shift all formants or only some of them (see Fig. 12). Again, this type of modification is almost trivial, because a simple multiplication of formant frequencies is sufficient. Experiment on unvoiced speech, showed that female voices can be transformed in order to give the impression of a male voice and vice versa, using simple formant shifts. In the case of unvoiced speech, the modification are very realistic, because no influence of glottal parameter have to be taken into account. Changes in formant bandwidth results also in changes of voice quality.

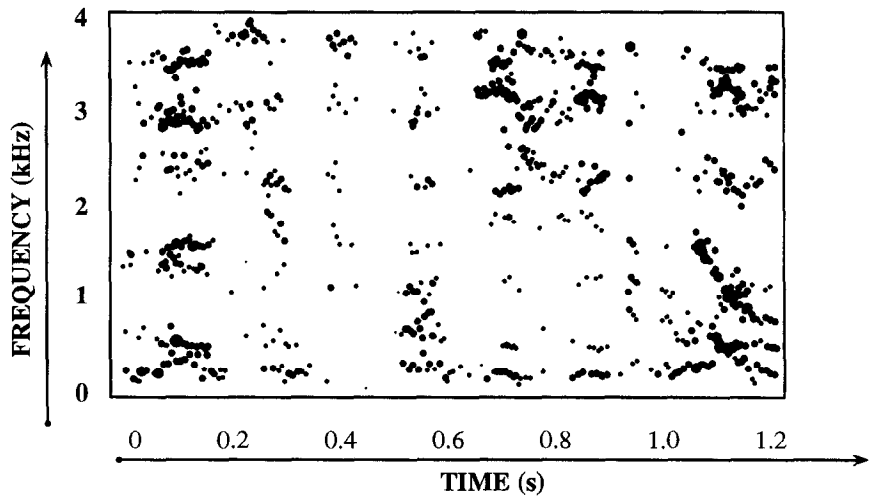

Fig. 12. At the end of the signal, the original formant rise seen in Fig. 10 is inversed. Audiofiles available (http://www.elsevier.nl/locate/specom).

6.3. Vocal effort modification

The FWF amplitude is another very important parameter since it allows to change voice characteristics such as spectral tilt and noise amplitude in selected regions.

Using the FWF bandwidth and excitation time parameters, it is possible to control the time domain envelope which characterizes the modulation structure of the noise. It is thus possible to modify the overall depth of the time modulation of the stochastic component. In a perceptual point of view, this is particularly important since deeper modulation gives a rougher voice with a impression of evident vocal effort, and a smoother modulation gives a softer and more whispery voice.

It is thus possible to modify the perceived vocal effort by joint modification of spectral tilt and temporal noise envelope. Furthermore, by performing a joint modification of the aperiodic component impulsiveness and periodic/aperiodic ratios, it is possible to change continuously from voiced to whispered speech which contains almost no modulation structure. With such modifications, voices with various degree of breathiness or creakiness can be synthesized.

7. Conclusion

Most studies in speech synthesis primarily focus on the quasi-periodic component of speech, neglecting the importance of the aperiodic component. In this paper, it is shown that a precise modelling of the aperiodic component is needed for voice quality speech modifications linked to noise (e.g. breathiness, creakiness, vocal effort). The aperiodic component is obtained by means of a recent algorithm which provides a meaningful decomposition of a speech signal into a periodic and an aperiodic component containing aspiration, frication and transient noises. The model proposed in this paper represents the aperiodic component as a sum of parameterized elementary waveforms well localized in the spectro-temporal domain. Despite a faithful representation of this component, this model provides original speech modification capabilities. Nevertheless further work must be devoted in this direction to establish a precise correlation between the different types of voice quality obtainable and the value of the synthesis parameters.

This work is based on Formant waveforms which have been chosen for their sucessful use in speech and singing synthesis. However, it appears that other types of waveforms might be more appropriate for random-like signals. A careful observation of the time domain envelope of the bandpass random signals analyzed shows that this envelope seems to be symmetric around local maxima (which is not the case for the FWF time domain envelope). This suggests that simple parameterized waveforms defined by sinusoids modulated by a symmetric time domain envelope (such as hamming window) could perform better data rate reduction with similar quality and speech modications abilities. However, in this case it would be more difficult to link those waveforms to speech perception or speech production considerations.

Acknowledgements

We would like to thank all persons who accept with kindness to participate to the perception tests. A special thanks goes to Yannis Stylianou for guidance in the implementation of the mLPC method, and to Wolfgang Minker for the abstract in German. We also are especially grateful to three anonymous reviewers for their helpful comments on an earlier version of this paper.

References

C. d'Alessandro (1989), Représentation du signal de parole par une somme de fonctions élémentaires, PhD dissertation, University of Paris VI, April 1989 (in French).

C. d'Alessandro (1990), "Time-frequency speech transformation based on an elementary waveform representation", *Speech Communication*, Vol. 9, Nos. 5/6, pp. 419–431.

C. d'Alessandro, B. Yegnanarayana and V. Darsinos (1995a), "Decomposition of speech signals into deterministic and stochastic components", *Proc. IEEE-ICASSP'84 Internat. Conf. Acoust. Speech Signal Process.*, Detroit, MI, pp. 760–763.

C. d'Alessandro, V. Darsinos and B. Yegnanarayana (1995b), "Evaluation of periodic/aperiodic decomposition for analysis of aperiodicities in the voice source", *Proc. ISMA'95 Internat. Symp. on Music. Acoust.*, Dourdan, France, pp. 446–452.

L.B. Almeida and F.M. Silva (1984), "Variable-frequency synthesis: An improved harmonic coding scheme", *Proc. IEEE-ICASSP'84 Internat. Conf. Acoust. Speech Signal Process.*, San Diego, CA, pp. 27.5.

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 Internat. Conf. Acoust. Speech Signal Process.*, Paris, France, pp. 614–617.

B.S. Atal and M.R. Schoeder (1970), "Adaptive predictive coding of speech signals", *Bell Syst. Tech. J.*, Vol. 49, pp. 1973–1986.

C. Berthomier (1983), "Instantaneous frequency and energy distribution of a signal", *Signal Processing.*, Vol. 5, No. 1, pp. 31–45.

H. Carl and B. Kolpatzik (1991), "Speech coding using nonstationary sinusoidal modelling and narrow-band basis functions", *Proc. IEEE-ICASSP'91 Internat. Conf. Acoust. Speech Signal Process.*, Toronto, Canada, pp. 581–584.

CCITT (1992), Revised recommendation P.80 – "Methods for subjective determination of transmission quality", SQEG, COM XII-118 E, Internat. Telegraph and Telephone Consultative Commitee (CCITT) from Recommendation P.80, Blue Book, Vol. V, 1989.

C. Chafe (1990), "Pulsed noise in self-sustained oscillations of musical instruments", *Proc. IEEE-ICASSP'90 Internat. Conf. Acoust. Speech Signal Process.*, Albuquerque, NM, pp. 1157–1160.

D.G. Childers and C.K. Lee (1991), "Vocal quality factors: Analysis, synthesis and perception", *J. Acoust. Soc. Amer.*, Vol. 90, No. 5, pp. 2394–2410.

W.B. Davenport Jr. (1952), "An experimental study of speech wave probability distributions", *J. Acoust. Soc. Amer.*, Vol. 24, No. 4, pp. 390–399.

W.B. Davenport Jr. and W.L. Root (1958), *An introduction to the Theory of Random Signals and Noise* (McGraw-Hill, New York), pp. 113–167.

G. De Krom (1993), "A cepstrum-based technique for determining a harmonics-to noise ratio in speech signals", *J. Speech Hearing Res.*, Vol. 36, pp. 254–266.

P. Depalle (1991), Analyse, Modèlisation et Synthèse des sons basées sur le modèle source-filtre, Ph.D. Dissertation, Université du Maine (in French).

T. Dutoit and H. Leich (1993), "MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database", *Speech Communication*, Vol. 13, Nos. 3–4, pp. 435–440.

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

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

J.L. Flanagan (1980), "Parametric coding of speech spectra", *J. Acoust. Soc. Amer.*, Vol. 68, pp. 412–419.

T. Foucart (1991), *Introduction aux Tests Statistiques* (Ed. Technip, Paris) (in French).

E.B. George and M.J.T. Smith (1992), "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones", *J. Audio Eng. Soc.*, Vol. 40, No. 6, pp 497–516.

I.R. Grandsten and S.W. Beet (1993), "Computationally efficient methods of calculating instantaneous frequency for auditory analysis", *Proc. Eurospeech'93 European Conf. on Speech Comm. and Tech.*, Berlin, pp. 385–389.

D. Griffin and J.S. Lim (1988), "Multiband excitation vocoder", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-36, No. 8, pp. 1223–1235.

D.J. Hermes (1991), "Synthesis of breathy vowels: Some research methods", *Speech Communication*, Vol. 10, Nos. 5–6, pp. 497–502.

J. Hillenbrand (1987), "A methodological study of perturbation and additive noise in synthetically generated voice signals", *J. Speech Hearing Res.*, Vol. 30, pp. 448–461.

N. Hiraoka, Y. Kitazoe, H. Ueta, S. Tanaka and M. Tanabe (1984), "Harmonic-intensity analysis of normal and hoarse voices", *J. Acoust. Soc. Amer.*, Vol. 76, pp. 1648–1651.

J.N. Holmes (1973), "The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer", *IEEE Trans. Audio Electroacoust.*, Vol. AU-21, pp. 298–305.

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

D.H. Klatt (1980), "Software for a cascade/parallel formant synthesizer", *J. Acoust. Soc. Amer.*, Vol. 67, No. 3, pp. 971–995.

D.H. Klatt and L.C. Klatt (1990), "Analysis, synthesis,and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Amer.*, Vol. 87, No. 2, pp. 820–857.

F. Klingholz (1987), "The measurement of the signal-to-noise ratio (SNR) in continuous speech", *Speech Communication*, Vol. 6, No. 1, pp. 15–26.

H. Kojima, W.J. Gould, A. Lambiase and N. Isshiki (1980), "Computer analysis of hoarseness", *Acta Otolaryngol.*, Vol. 89, pp. 547–554.

J. Laroche, Y. Stylianou and E. Moulines (1993), "HNS: Speech modification based on a harmonic + noise model", *Proc. IEEE-ICASSP'93 Internat. Conf. Acoust. Speech Signal Process.*, Minneapolis, MN, pp. 550–553.

M.R. Leadbetter (1972), "Point processes generated by level crossings", in: P.A.W. Lewis, Ed., *Stochastic Point Processes, Statistical Analysis, Theory and Applications* (Wiley, New York), pp. 436–467.

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

J.C.W.A. Liljencrants (1968), "The OVE III speech synthesizer", *IEEE Trans. Audio Electroacoust.*, Vol. AU-16, No. 1, pp. 137–140.

L. Mandel (1967), "Complex representation of optical fields in coherence theory", *J. Opt. Soc. Amer.*, Vol. 57, No. 5., pp. 613–617.

J.S. Marques and A.J. Abrantes (1994), "Hybrid harmonic coding of speech at low bit rates", *Speech Comunication*, Vol. 14, No. 3, pp. 231–247.

R.J. McAulay and T.F. Quatieri (1992), "Low-rate speech coding based on the sinusoidal model", in: S. Furui and M. M. Sondhi, Eds., *Advances in Speech Signal Processing* (Marcel Dekker, New York), pp. 165–208.

E. Moulines and F. Charpentier (1990), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol. 9, Nos. 5/6, pp. 453–467.

M.D. Paez and T.H. Glisson (1972), "Minimum mean squared-error quantization in speech", *IEEE Trans. Comm.*, Vol. Com-20, pp. 225–230.

A. Papoulis (1983), "Random modulation: a review", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-31, No. 1, pp. 96–105.

A. Papoulis (1986), *Probability, Random Variables, and Stochastic Processes*, 2nd edition (McGraw-Hill, New York).

B. Picinbono (1989), "The analytical signal and related problem", in: G. Longo and B. Picinbono. Eds., *Time and Frequency Representation of Signals and Systems*, CISM Courses and Lectures No. 309 (Springer, Wien), pp. 1–9.

B. Picinbono and W. Martin (1983), "Représentation des signaux par amplitude et phase instantanées", *Ann. Télécommun.*, Vol. 38, No. 5–6, pp. 179–190 (in French).

L.R. Rabiner (1968), "Digital-formant synthesizer for speech-synthesis studies", *J. Acoust. Soc. Amer.*, Vol. 49, No. 4, pp. 822–828.

S.O. Rice (1944–1945), "Mathematical analysis of random noise", *Bell Syst. Tech. J.*, Vol. 24, pp. 282–332, Vol. 25, pp. 46–156.

G. Richard (1994), Modèlisation de la composante stochastique de la parole, PhD thesis, Université de Paris-XI, Orsay, France (in French).

G. Richard, C. d'Alessandro and S. Grau (1992), "Unvoiced speech analysis and synthesis using Poissonian random formant-wave-functions", *Proc. EUSIPCO'92 European Sig. Process. Conf.*, Brussels, Belgium, pp. 347–350.

G. Richard, C. d'Alessandro and S. Grau (1993), "Musical noises synthesis using random waveforms", *Proc. SMAC'93 Stock. Music. Acoust. Conf.*, Stockholm, Sweden, pp. 580–583.

X. Rodet (1980), "Time-domain formant-wave-function synthesis", in: J.C. Simon, Ed., *Spoken Language Generation and Understanding* (Reidel, Dordrecht). Reprinted in *Computer Music J.*, Vol. 8, No. 3, pp. 9–14.

X. Rodet, P. Depalle and G. Poirot (1987), "Speech analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions", *Proc. European Conf. on Speech Comm. and Tech.*, Edinburgh, UK.

X. Serra and J. Smith (1990), "Spectral modeling synthesis: A sound/synthesis system based on a deterministic plus stochastic decomposition", *Computer Music J.*, Vol. 14, No. 4.

D.L. Snyder (1975), *Random Point Processes* (Wiley/Interscience, New York).

K.N. Stevens (1971), "Airflow and turbulence noise for fricative and stop consonants: static considerations", *J. Acoust. Soc. Amer.*, Vol. 50, No. 2, pp. 1180–1192.

P. Strevens (1960), "Spectra of fricative noise in human speech", *Language and Speech*, Vol. 3. Reprinted in: Lehiste, Ed., *Readings in Acoustic Phonetics* (MIT press, Cambridge, MA, 1967), pp. 202–219.

A. Tsopanoglou, J. Mourjopoulos and G. Kokkinakis (1993), "Speech representation and analysis by the use of instantaneous frequency", in: M. Cooke, S. Beet and M. Crawford, *Visual Representation of Speech Signals* (Wiley, New York), pp. 341–346.

J. Wendler, H. Wagner, W. Seidner and E. Stuerzebecher (1976), "Methodik bei Stimmschalanalysen", *16ième Cong. Logopedics and Phoniatrics*, Interlaken, 1974, Conference Record (Kargel, Basel), pp. 518–521.

E. Yumoto, W.J. Gould and T. Baer (1982), "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Amer.*, Vol. 71, pp. 1544–1550.