Non-parametric technique for pitch-scale and time-scale modification of speech

Article *in* Speech Communication · February 1995

Impact Factor: 1.26 · DOI: 10.1016/0167-6393(94)00054-E · Source: DBLP

2 authors, including:

Eric Moulines
MINES ParisTech

331 PUBLICATIONS **13,543** CITATIONS

SEE PROFILE

ELSEVIER

Speech Communication 16 (1995) 175–205

SPEECH COMMUNICATION

Non-parametric techniques for pitch-scale and time-scale modification of speech

Eric Moulines *, Jean Laroche

Télécom Paris, 46 Rue Barrault, 75634 Paris Cedex 13, France

Received 29 March 1994; revised 27 October 1994

Abstract

Time-scale and, to a lesser extent, pitch-scale modifications of speech and audio signals are the subject of major theoretical and practical interest. Applications are numerous, including, to name but a few, text-to-speech synthesis (based on acoustical unit concatenation), transformation of voice characteristics, foreign language learning but also audio monitoring or film/soundtrack post-synchronization. To fulfill the need for high-quality time and pitch-scaling, a number of algorithms have been proposed recently, along with their real-time implementation, sometimes for very inexpensive hardware. It appears that most of these algorithms can be viewed as slight variations of a small number of basic schemes. This contribution reviews frequency-domain algorithms (phase-vocoder) and time-domain algorithms (Time-Domain Pitch-Synchronous Overlap/Add and the like) in the same framework. More recent variations of these schemes are also presented.

Zusammenfassung

Modifikationen von Sprachsignalen und akustischen Signalen im Zeitbereich und, in geringerem Maße, im Grundfrequenzbereich, sind von großer theoretischer und praktischer Bedeutung. Es gibt zahlreiche Anwendung, wie z.B. die Sprachvollsynthese (durch Verkettung von akustischen Einheiten), die Umwandlung von Sprachmerkmalen, Fremdsprachenerwerb, aber auch Audiokontrolle und Postsynchronisierung. Um den Anforderungen nach hoher Qualität gerecht zu werden, wurden in letzter Zeit eine gewisse Anzahl an Algorithmen mit ihrer Echtzeitimplementierung vorgestellt, manchmal für sehr kostengünstige Hardwares. Es kann davon ausgegangen werden, daß die meisten dieser Algorithmen nur geringfügige Varianten desselben Grundschemas sind. Dieser Artikel gibt einen Überblick über die Algorithmen im Frequenzbereich (phase-vocoder) und im Zeitbereich (Time-Domain Pitch-Synchronous Overlap/Add), die sich im gleichen Rahmen bewegen. Mehrere neuere Methoden, die aus diesen Techniken hervorgegangen sind, werden ebenfalls beschrieben.

Résumé

Les techniques de modifications de l'échelle temporelle et, à un degré moindre, de la fréquence fondamentale présentent un intérêt théorique et pratique certain. Les applications sont nombreuses, et incluent, pour n'en citer que quelques-unes, la synthèse de parole à partir du texte (par concaténation d'unités acoustiques), les transformations du timbre de la voix, l'apprentissage des langues étrangères, mais aussi le contrôle audio ou la post-synchronisation. Pour satisfaire ces besoins, un grand nombre d'algorithmes ont été proposés ces dernières années,

* Corresponding author.

SSDI 0167-6393(94)00054-9

apparemment différents mais en fait très proches sur le plan des mécanismes de base. Cet article présente les méthodes frequentielles (vocodeur de phase) et temporelles (Time-Domain Pitch-Synchronous Overlap/Add) dans un même formalisme. Plusieurs méthodes plus récentes, dérivées de ces techniques fondamentales sont également décrites.

Keywords: Pitch-scale and time-scale transformations; Phase vocoder; PSOLA analysis–synthesis; Quasi-harmonic model

1. Introduction

Time-scale and, to a lesser extent, pitch-scale modifications of speech and audio signals are the subject of major theoretical and practical interest. Applications are numerous, including, to name but a few, text-to-speech synthesis (based on acoustical unit concatenation), transformation of voice characteristics, foreign language learning but also audio monitoring or film/soundtrack post-synchronization.

To fulfill the need for high-quality time and pitch-scaling, a number of algorithms have been proposed recently, along with their real-time implementation, sometimes for very inexpensive hardware. It appears that most of these algorithms can be viewed as slight variations of a small number of basic schemes.

It is the main purpose of this contribution to review these algorithms in the same common framework, based on a simple extension of the short-time Fourier transform analysis–synthesis principle allowing time-varying analysis/synthesis rates.

The paper is organized as follows. In Section 2, a basic speech production model is introduced, that helps define precisely time-scale and pitch-scale modifications. The short-time Fourier transform is then briefly presented. In Section 3, time-scale and pitch-scale speech modifications are addressed, based on the standard, uniform rate STFT analysis/synthesis scheme (referred to as the *phase vocoder* to stick with the tradition of the speech and audio community). Much of the material in this section is inspired by the pioneering contributions of M. Portnoff for time-scale modification (Portnoff, 1981b) and S. Seneff for pitch-scale modification (Seneff, 1982). In Section 4, the Pitch-Synchronous Overlap-Add (PSOLA) synthesis technique is presented. Links with the phase-vocoder techniques are evidenced. Several variations of the basic PSOLA scheme are also documented. Section 5 presents recent contributions and variations on these algorithms.

2. Preliminaries

2.1. The short-time Fourier Transform and the phase vocoder

The short-time Fourier Transform (STFT) methods have been used for speech analysis, synthesis and modifications for many years and the applications are numerous (early implementations of the so-called phase vocoder date back as early as 1966 (Flanagan and Golden, 1966)). Its theory is now well understood and efficient implementations are available, based on FFT and simple modifications of Overlap and Add synthesis methods. Most of the modern results can be found in the papers by Crochiere (Crochiere, 1980, Crochiere and Rabiner, 1983) which clarify the links between filter bank summation techniques and block implementations of the STFT; see (Allen, 1982) for an introduction to these methods; see also (Nawab and Quatieri, 1988) for more recent references.

In this section, we first present the basic mechanisms behind the short-time Fourier transform analysis/synthesis. The presentation is slightly more general than the traditional derivations, to allow variable analysis and synthesis rates, which are useful in certain applications.

Analysis. The short-time Fourier transform can be viewed as an alternate way of representing a signal in a joint time and frequency domain. The basic idea consists of performing a Fourier transform on a limited portion of the signal, then shifting to another portion of the signal and re-

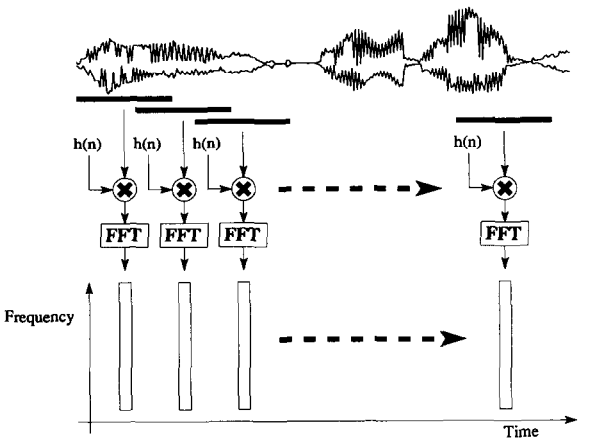

Fig. 1. The short-time Fourier Transform.

peating the operation, as shown in Fig. 1. The signal is then described by the values of the Fourier transforms obtained at the different window locations.

Henceforth, we denote the sampled signal to be analyzed by $x(n)$ (the sampling period is assumed to be equal to $T_e = 1$). The successive window locations, referred to as analysis time-instants are denoted $t_a(u)$ (they are assumed to take only integer values). In standard applications, the STFT analysis is performed at a constant rate: the analysis time-instants are regularly spaced, i.e. $t_a(u) = uR$, where R is a fixed integer increment which controls the analysis rate. As outlined above, non-uniform analysis rates are sometimes required, essentially for time-scale and pitch-scale transformation: they are exploited for example in the Pitch Synchronous Overlap Add method (PSOLA) (see Section 4).

Formally, the short-time Fourier transform can be described in either a low-pass or a band-pass convention. We will only describe the band-pass convention, since it corresponds to the way the transform is carried out in practice. The low-pass convention is more useful for theoretical analysis, see for example (Portnoff, 1981a). In the band-pass convention, the short-time Fourier transform $X(t_a(u),\omega)$ is given by

$$X(t_a(u),\omega) = \sum_{n=-\infty}^{\infty} h_u(n)x(t_a(u)+n)\exp(-j\omega n). \quad (1)$$

The analysis window $h_u(n)$ is a weighting function which is applied to the input signal prior to the Fourier transform (it selects and weights the short-time segment of the signal to be analyzed). $X(t_a(u),\omega)$ is *the short-time analysis spectrum* of the signal at time $t_a(u)$: it is interpreted, according to the above expression, as the discrete Fourier transform of the *analysis short-time signal*

$$x(t_a(u),n) = h_u(n)x(t_a(u)+n). \quad (2)$$

In traditional presentations (constant analysis rate), the analysis window is a fixed window function, i.e. $h_u(n) = h(n)$ for every analysis time-instants $t_a(u)$. In the more general framework used here (time-varying analysis rate), the analysis window may depend on the analysis time-index u, hence the additional subscript u.

Implicitely, the analysis window is (i) of finite duration T_u and symmetric (i.e. non-zero between $[-T_u/2, T_u/2]$,) and (ii) of low-pass type, i.e. $h_u(n)$ is the impulse response of a low-pass filter, with a small time-bandwidth product γ. [1] Any of the numerous windows proposed in the literature can be used, depending on their spectral characteristics (mainlobe bandwidth and side-lobe attenuation): possible choices include Hanning, Blackman, Kaiser,... When the analysis window is kept constant, an alternative solution consists of designing a symmetric low-pass FIR filter with an appropriate cutoff frequency by use of any FIR filter design technique.

When $t_a(u) = uR$ and $h_u(n) = h(n)$ (i.e. the analysis rate and the analysis window are constant), it is easy to see that Eq. (1) corresponds to the convolution of $x(n)$ by a filter whose impulse response is $h(-n)\exp(j\omega n)$. Since $h(-n)$ is the impulse response of a low-pass filter, $h(-n)\exp(j\omega n)$ is the impulse response of a band-pass filter centered at frequency ω.

As a result, for a given value of ω, the short-time Fourier transform $X(t_a(u),\omega)$ can be viewed as the output of a complex band-pass filter cen-

[1] The time-bandwidth product is defined for low-pass filters as the product of the length of the filter impulse response and twice its cutoff frequency; it is a dimensionless number which quantifies the "quality" of the window.

tered at ω and sampled at the successive time-instants $t_a(u)$.

In practice, the STFT is evaluated at discrete frequencies $\Omega_k = 2\pi k/N$ by use of a Fast Fourier Transform (FFT) of length N. Note that N must be larger than the length T_u of the analysis window $h_u(n)$ for any value of the time index u. In audio and speech processing, the term *phase vocoder* refers to STFT analysis–synthesis, in which the complex values of the short-time spectrum components $X(t_a(u),\Omega_k)$ are expressed in polar coordinates, magnitude M and phase ϕ (hence the name):

$$X(t_a(u),\Omega_k) = M(t_a(u),\Omega_k)\exp(j\phi(t_a(u),\Omega_k)) \stackrel{\text{def}}{=} M_k(u)\exp(j\phi_k(u)). \quad (3)$$

Modification and Synthesis. The discussion up to this point has been centered around analysis. We now move to the modification and synthesis stages, which are of fundamental importance for speech processing applications.

Modification stage: For time-scale and pitch-scale applications, the modification stage consists of (i) applying appropriate modifications to the stream of short-time analysis spectra $X(t_a(u),\Omega_k)$ to produce a stream of *short-time synthesis spectra* $Y(t_s(u),\Omega_k)$ and (ii) synchronizing these short-time synthesis spectra $Y(t_s(u),\Omega_k)$ on a new set of time-instants, referred to as the synthesis time-instants and denoted $t_s(u)$. Examples of modification of the short-time spectrum include phase transformation (time-scaling) and frequency-axis compression/expansion (pitch-scaling), which are both detailed below. This list is obviously not exhaustive. The stream of synthesis time-instants $t_s(u)$ is determined from the stream of analysis time-instants according to the desired pitch-scale and time-scale modifications. The number of synthesis time-instants needs not be identical to the number of analysis time-instants (it is identical for phase vocoder transformations, but is not for PSOLA). For non-constant pitch-scale and time-scale modification factors, the synthesis time-instants will be generally irregularly spaced, whether or not the analysis rate is constant.

Synthesis stage: The last step consists of combining the stream of synthesis short-time signals synchronized on the synthesis time-instants to obtain the desired "modified" signal, as shown below. Given an arbitrary sequence of synthesis short-time Fourier transforms $Y(t_s(u),\Omega_k)$, there is in general no time-domain signal $y(n)$ of which $Y(t_s(u),\Omega_k)$ is the short-time Fourier transform: the stream of short-time Fourier transforms of a given signal must satisfy "strong" consistency conditions since the Fourier transforms correspond to *overlapping* short-time signals (these conditions are given for example in (Portnoff, 1980)). Several solutions to this problem exist. From a theoretical point of view, there are strong motivations towards using the weighted least-square overlap-add procedure (Griffin and Lim, 1984). This procedure consists of seeking the synthetic signal $y(n)$ whose short-time Fourier transform (around time instants $t_s(u)$)

$$\hat{Y}(t_s(u),\Omega_k) = \sum_m f_u(m)\, y(t_s(u)+m)\exp(-j\Omega_k m) \quad (4)$$

best fits the modified synthesis short-time Fourier transform $Y(t_s(u),\Omega_k)$, in the least-square sense (standard Euclidian norm)

$$\sum_u \sum_{k=0}^{N-1} |Y(t_s(u),\Omega_k) - \hat{Y}(t_s(u),\Omega_k)|^2; \quad (5)$$

$f_u(n)$, which is used in the definition of the short-time spectrum (Eq. (4)), is referred to as the *synthesis window*. In practical realizations, $f_u(n)$ is of finite-duration; to avoid time-aliasing, its duration is smaller than N, the number of FFT coefficients. It is usually directly deduced from the analysis window, reflecting the modifications brought to the synthesis short-time spectrum $Y(t_s(u),\Omega_k)$, on which $\hat{Y}(t_s(u),\Omega_k)$ is fitted. When the synthesis rate is constant, the synthesis window is a fixed window function, i.e. $f_u(n) = f(n)$, for all the synthesis indexes u. More generally, $f_u(n)$ can be time-dependent, hence the subscript u. Applying the Parseval relation, and under the assumption that the synthesis window length is less than the length of the transform, the minimization problem Eq. (5) may be ex-

pressed in the time-domain: the synthetic signal $y(n)$ must minimize the following least-square error $\mathscr{E}$:

$$\mathscr{E} = \sum_u \sum_n \left[y_w(u,n) - f_u(n)\, y(t_s(u)+n) \right]^2, \quad (6)$$

with

$$y_w(u,n) = \frac{1}{N} \sum_{k=0}^{N-1} Y(t_s(u),\Omega_k) \exp(j\Omega_k n). \quad (7)$$

A slight generalization of the above approach can be developed by weighting the least-square criterion by a sequence $w_u(n)$ of positive weighting functions.

$$\mathscr{E} = \sum_u \sum_n w_u(n) \times \left[y_w(u,n) - f_u(n)\, y(t_s(u)+n) \right]^2. \quad (8)$$

The least-square problem Eq. (8) can be solved explicitly; the solution is given in closed form by the following *synthesis formula*:

$$y(n) = \frac{\sum_u y_w(u,n-t_s(u)) f_u(n-t_s(u)) w_u(n-t_s(u))}{\sum_u w_u(n-t_s(u)) f_u^2(n-t_s(u))}. \quad (9)$$

The synthesis algorithm is similar to weighted overlap-add: the successive short-time synthesis signals are combined, with appropriate weights and time-shifts. [2] The denominator plays the role of a time-varying normalization factor, which compensates for the energy modifications resulting from the possibly variable overlap between the successive windows.

In absence of modifications, and by taking identical analysis and synthesis windows, i.e. $h_u(n) = f_u(n)$ for all u, the signal $y(n)$ resulting from the synthesis scheme Eq. (9) is exactly equal to the original signal $x(n)$, if

$$\sum_u w_u(n-t_s(u)) h_u^2(n-t_s(u)) \neq 0 \quad \forall n, \quad (10)$$

i.e. the weighted-synthesis windows do overlap. Under this condition (which is assumed to be fulfilled in the sequel), the STFT analysis–synthesis system achieves perfect reconstruction (more general conditions of perfect reconstruction involving different analysis/synthesis windows are discussed in (Crochiere and Rabiner, 1983)).

Different choices for the weighting function $w_u(n)$ yield different synthesis schemes. Maybe the most straightforward approach consists of setting $w_u(n) = 1$, leading to the least-square overlap/add procedure, initially proposed by Griffin and Lim (1984) (for fixed analysis and synthesis rate). In this case,

$$y(n) = \frac{\Sigma_u f_u(n-t_s(u)) y_w(u,n-t_s(u))}{\Sigma_u f_u^2(n-t_s(u))}. \quad (11)$$

Because of the particular form of the LS-criterion, this choice mainly weights the errors in the center of the synthesis window. To counter this effect, an appropriate choice is $w_u(n) = 1/f_u(n)$, for $f_u(n) \neq 0$; $w_u(n) = 0$, otherwise. For this particular choice, the synthesis formula becomes

$$y(n) = \frac{\Sigma_u y_w(u,n-t_s(u))}{\Sigma_u f_u(n-t_s(u))}, \quad (12)$$

which is nothing but the simple overlap-add synthesis procedure (see for example (Allen, 1977)).

In real-time implementations, the time-varying normalization factor $\Sigma_u f_u(n-t_s(u))$ appearing in the synthesis formula Eq. (12) may be undesirable because it requires one division per output sample. For constant rate analysis/synthesis, however, this normalization factor can be calculated once and for all, and included in the synthesis window. For moderate time-varying time/pitch modification rates, the normalization can generally be omitted without affecting the quality of the reconstructed output: when the analysis window is long enough compared to the spacing between analysis instants, this factor is indeed almost constant, a consequence of the low-pass characteristics of the analysis and synthesis windows.

[2] Time-shifts in the short-time signals only reflect the fact that we use a different time-origin for each short-time signal, as a consequence of the use of the band-pass convention; an equivalent low-pass formula may be derived, which assumes a time-origin common to all short-time synthesis signals.

2.2. Quasi-stationary representation of speech signal

When designing time-scale and pitch-scale modification algorithms, it is often convenient to refer to a parametric model for the production of speech signals (even when this model is not used explicitly for analysis/synthesis purposes). In this section, a flexible quasi-stationary representation of the sampled speech waveform is introduced (Portnoff, 1981a; McAulay and Quatieri, 1986). The relationship between the short-time Fourier transform of speech signal and the parameters of the underlying speech production model is then evidenced.

2.2.1. Speech production model

According to the generally accepted engineering model for speech production, the sampled speech waveform is modelled as the output of a time-varying linear filter driven by an excitation signal, which is either a sum of narrow-band signals with harmonically related instantaneous frequencies (voiced speech) or a stationary random sequence, with a flat power spectrum (unvoiced speech).

We will mainly focus on voiced speech. The time varying filter approximates the combined effect of (i) the transmission characteristics of the supra-glottal cavities (including the radiation at the mouth opening) and (ii) the glottal pulse shape. The input–output behavior of this system is characterized by its time-varying unit-sample response $g_n(m)$, defined as the response of the system at time n to a unit-sample applied at time $n-m$. An equivalent description is given by the time-varying transfer function of the system, defined as the Fourier transform of $g_n(m)$ with respect to index m,

$$\sum_{m=-\infty}^{+\infty} g_n(m)\exp(-j\omega m) = G(n,\omega)\exp(j\psi(n,\omega)), \tag{13}$$

$G(n,\omega)$ and $\psi(n,\omega)$ are respectively referred to as the time-varying amplitude and phase of the system. The non-stationarity of $g_n(m)$ corresponds to movements of physical articulators and is usually relatively slow compared to the time-variation of the speech waveform. It can be considered as nearly constant for the duration of its memory, i.e. $g_n(m)$ is a *quasi-stationary* system. For voiced speech, the excitation waveform $e(n)$ is represented as a sum of harmonically related complex exponentials with unit amplitudes, zero initial phase, and a slowly varying fundamental frequency function $n \to 2\pi/P(n)$, where $P(n)$ is the local pitch-period (the function $n \to P(n)$ is referred to as the *pitch contour*). In mathematical terms, this is expressed as

$$e(n) = \sum_{k=0}^{P(n)-1} \exp[j(\Phi_k(n))], \tag{14}$$

where $\Phi_k(n)$ is the excitation phase of the k-th harmonic: it is defined as the integral of the time-varying harmonic frequency $\omega_k(n) = 2\pi k/P(n)$,

$$\Phi_k(n) = \sum_{m=0}^{n} \omega_k(m) = \sum_{m=0}^{n} \frac{2\pi k}{P(m)}. \tag{15}$$

Note that the amplitudes of the pitch-harmonics in the excitation signal have been assumed constant: $G(n,\omega)$ alone accounts for the magnitude of the speech signal's spectrum. Similarly, the pitch-harmonics in the excitation signal have a null initial phase: the system phase $\psi(n,\omega)$ alone accounts for the phases of the signal's pitch-harmonics.

Because $P(n)$ is nearly constant around the time instant n, the excitation phase $m \to \Phi_k(m)$ may be approximated, in the neighborhood of n as

$$\Phi_k(m) \approx \Phi_k(n) + \omega_k(n)(m-n) \quad \text{for small } |m-n|. \tag{16}$$

According to standard time-varying filtering formulas, the voiced speech signal $x(n)$ modeled as the output of $g_n(m)$ driven by $e(m)$ is given by

$$x(n) = \sum_{m=-\infty}^{+\infty} g_n(m)e(n-m). \tag{17}$$

Assuming that the pitch-period $P(n)$ is constant for the duration of $g_n(m)$, the excitation

signal can be replaced by its local harmonic representation to obtain

$$x(n) = \sum_{k=0}^{P(n)-1} G(n,\omega_k(n)) \times \exp\left[j(\Phi_k(n) + \psi(n,\omega_k(n)))\right] = \sum_{k=0}^{P(n)-1} A_k(n)\exp(j\theta_k(n)). \quad (18)$$

The harmonic amplitude $A_k(n)$ of the k-th harmonic is the system amplitude $A_k(n) = G(n,\omega_k(n))$ at the harmonic frequency $\omega_k(n)$: The phase $\theta_k(n)$ of the k-th harmonic is the sum of the excitation phase $\Phi_k(n)$ and the system phase $\psi_k(n) = \psi(n,\omega_k(n))$:

$$\theta_k(n) = \Phi_k(n) + \psi(n,\omega_k(n)) = \Phi_k(n) + \psi_k(n). \quad (19)$$

$\theta_k(n)$ is often referred to as the *instantaneous phase* of the k-th harmonic. The system phase $\psi(n,\omega_k(n))$ being a slowly varying function of n, the instantaneous phase $m \to \theta_k(m)$ may be developed, according to Eq. (16), in the neighborhood of n as

$$\theta_k(m) = \theta_k(n) + \omega_k(n)(m-n) \quad \text{for small } |m-n|. \quad (20)$$

Further discussion on this model and its validity can be found in (Portnoff, 1981a).

Short-time Fourier analysis of voiced speech

The short-time Fourier transform of a voiced speech signal $x(n)$ may be easily expressed in terms of its harmonic representation. Substituting the harmonic representation Eq. (18) for $x(n)$ into Eq. (1) gives

$$X(t_a(u),\Omega_l) = \sum_{m=-\infty}^{+\infty} h_u(m) \sum_{k=0}^{P(m)-1} A_k(t_a(u)+m) \times \exp\left[j\theta_k(t_a(u)+m)\right]\exp(-j\Omega_l m) \quad (21)$$

for $0 \leqslant l < N$. We now assume that the duration of the analysis window $h_u(m)$ is sufficiently short so that the pitch period $m \to P(t_a(u)+m)$ and the harmonic amplitudes $m \to A_k(t_a(u)+m)$ are constant over the duration of $h_u(m)$. Assuming that the local representation Eq. (20) of the instantaneous phase $\theta_k(t_a(u)+m)$ holds,

$$\theta_k(t_a(u)+m) = \theta_k(t_a(u)) + m\omega_k(t_a(u)),$$

it can be used in Eq. (21), giving

$$X(t_a(u),\Omega_l) = \sum_{k=0}^{P(t_a(u))-1} A_k(t_a(u))\exp(j\theta_k(t_a(u))) \times H_u(\Omega_l - \omega_k(t_a(u))), \quad (22)$$

where $H_u(\omega)$ is the DFT of the analysis window $h_u(n)$. Eq. (22) expresses the STFT $X(t_a(u), \Omega_l)$ as the sum of $P(t_a(u))$ images of $H_u(\omega)$ each shifted in frequency by $\omega_k(t_a(u))$ and weighted by $A_k(t_a(u))\exp(j\theta_k(t_a(u)))$.

From now on, the analysis window is assumed to be symmetric around zero: as a consequence, $H_u(\omega)$ is real. Also, for phase-vocoder based time-scale and pitch-scale transformations (see below), the cutoff frequency ω_h of the analysis window $h_u(m)$ is chosen to be less than half the spacing between the pitch-harmonics: this is referred to as the *narrow-band analysis condition* (Portnoff, 1981a). [3] The shifted and weighted images of the window $H_u(\omega)$ are therefore non-overlapping, and $X(t_a(u), \Omega_l)$ reduces to

$$X(t_a(u),\Omega_l) = \begin{cases} A_k(t_a(u))\exp[j\theta_k(t_a(u))] \times H_u(\Omega_l - \omega_k(t_a(u))) & \text{for } |\Omega_l - \omega_k(t_a(u))| \leqslant \omega_h, \\ 0 & \text{otherwise.} \end{cases} \quad (23)$$

The magnitude $M(t_a(u), \Omega_l)$ of the STFT,

$$M(t_a(u),\Omega_l) \overset{\text{def}}{=} M_l(u) = A_k(t_a(u))H_u(\Omega_l - \omega_k(t_a(u))) \quad \text{for } |\Omega_l - \omega_k(t_a(u))| \leqslant \omega_h, \quad (24)$$

is a slowly varying function of $t_a(u)$ because both $A_k(t_a(u))$ and $P(t_a(u))$ are slowly varying func-

[3] Of course, the analysis windows are not strictly band-limited; ω_h is the width of the window's mainlobe.

tions of $t_a(u)$. Similarly, the phase $\phi(t_a(u),\Omega_l)$ of $X(t_a(u),\Omega_l)$ may be expressed as

$$\phi(t_a(u),\Omega_l) \stackrel{\text{def}}{=} \phi_l(u) = \arctan\left(\frac{\Im(X(t_a(u),\Omega_l))}{\Re(X(t_a(u),\Omega_l))}\right) \equiv \theta_k(t_a(u)) \quad \text{for } |\Omega_l - \omega_k(t_a(u))| \leqslant \omega_h, \quad (25)$$

where $a \equiv b$ indicates that $a = b \bmod 2\pi$. It is important to note that only the principal determination of the phase is accessible using the short-time Fourier analysis.

Instantaneous frequency. The instantaneous phase $\phi_l(u)$ carries information on the *instantaneous frequency* $\omega_k(t_a(u))$ of the single pitch-harmonic that "falls" into the l-th channel. Provided that the successive time-instants $t_a(u)$ and $t_a(u-1)$ are close enough so that the local expansion of the instantaneous phase given in Eq. (20) applies, the instantaneous frequency $\omega_k(t_a(u))$ can be determined, up to an integer multiple of 2π, by computing the first-order backward difference $\Delta\phi_l(u)$ of the instantaneous phase $\phi_l(u)$:

$$\Delta\phi_l(u) \stackrel{\text{def}}{=} \phi(t_a(u),\Omega_l) - \phi(t_a(u-1),\Omega_l) = (t_a(u) - t_a(u-1))\omega_k(t_a(u)) + 2n\pi. \quad (26)$$

The number n is unknown, but can be estimated by "unwrapping" the phase in the time-domain as is described below. The preceding equation can be rewritten as

$$\Delta\phi_l(u) = \Omega_l(t_a(u) - t_a(u-1)) + (\omega_k(t_a(u)) - \Omega_l) \times (t_a(u) - t_a(u-1)) + 2n\pi. \quad (27)$$

We will denote $R(u) = t_a(u) - t_a(u-1)$ the interval in samples between the two successive analysis instants. Assuming the k-th pitch-harmonic falls within the l-th channel, we have

$$|(\omega_k(t_a(u)) - \Omega_l)R(u)| < \omega_h R(u), \quad (28)$$

where ω_h is the bandwidth of the analysis window. Assuming that $R(u)$ is such that $\omega_h R(u) < \pi$, and using Eqs. (27) and (28) we find

$$|\Delta\phi_l(u) - \Omega_l R(u) - 2n\pi| < \pi.$$

There is only one integer n that satisfies the latter inequality, and it can be determined from the value of $\Delta\phi_l(u) - \Omega_l R(u)$. Once the value of m is determined, the instantaneous frequency is derived from Eq. (27).

The calculation of the instantaneous frequency in the channel can be summarized as follows:

1. From two successive short-time spectra $X(t_a(u-1),\omega_k)$ and $X(t_a(u),\Omega_l)$, calculate the phase increment $Z(t_a(u),\Omega_l) = \phi(t_a(u),\Omega_l) - \phi(t_a(u-1),\Omega_l) - \Omega_l(t_a(u) - t_a(u-1))$.
2. Modify the phase increment by adding or subtracting multiples of 2π so the result $\bar{Z}(t_a(u),\Omega_l)$ lies between $-\pi$ and π.
3. Estimate the instantaneous frequency $\lambda(t_a(u),\Omega_l)$ in the l-th channel by

$$\lambda(t_a(u),\Omega_l) = \Omega_l + \frac{\bar{Z}(t_a(u),\Omega_l)}{t_a(u) - t_a(u-1)}. \quad (29)$$

As outlined above, in the present model the instantaneous frequency $\lambda(t_a(u), \Omega_l)$ in the l-th channel is the instantaneous frequency of the pitch-harmonic that falls into the channel: $\lambda(t_a(u), \Omega_l) = \omega_k(t_a(u))$. The definition and the estimation of the instantaneous frequency in each STFT channel remains valid in more general conditions; in fact, it is sufficient that at most one sinusoidal component be present in each channel.

Note that although the operation that converts the real and imaginary parts of $X(t_a(u),\Omega_l)$ into values of amplitude and instantaneous frequency is non-linear, it can easily be inverted, i.e., the real and imaginary parts of $X(t_a(u),\Omega_l)$ can be calculated from the successive values of $M(t_a(u),\Omega_l)$ and $\bar{Z}(t_a(u),\Omega_l)$, up to an initial phase.

Choice of the analysis parameters. We can now summarize the various constraints introduced so far on the length of the analysis window T_u, its cutoff frequency ω_h, and the analysis rate R:

- For the STFT analysis to resolve the pitch-harmonics, the cutoff frequency of the analysis window must satisfy $\omega_h < \pi/P(t_a(u))$, i.e. be less than half the spacing between two successive pitch-harmonics (condition of narrow-band analysis).
- The duration of the analysis window T_u must

be small enough so the amplitudes and instantaneous frequencies of the pitch-harmonics can be considered constant within the analysis window.
- The FFT size N must be larger than the window length $N > T_u$.
- To make phase unwrapping possible, the cutoff frequency and the analysis rate must satisfy $\omega_h R < \pi$.

For standard analysis windows (e.g. Hanning, Hamming) the cutoff frequency is inversely proportional to the window length, $\omega_h = 4\pi/T_u$. The first condition (narrow band analysis) thus implies that $T_u > 4P(t_a(u))$, the local pitch-period: *The window must be longer than 4 pitch-periods*. To make the STFT analysis/synthesis system more robust to speaker or speaking-conditions variability, the window length can be adapted to the average pitch-period.

The last constraint above implies $R < T_u/4$, i.e. successive analysis windows must have a minimum overlap of 75%. The larger the cutoff frequency, the larger the minimum overlap between successive analysis windows.

2.3. Time / pitch scale specification

In this section, ideal time-scale and pitch-scale modifications are addressed with reference to the speech production model outlined above. We first define the time/pitch-scale warping function, then describe the effect of time/pitch scale modifications on the model parameters.

2.3.1. Time-scale modification

The object of time-scale modification is to alter the apparent rate of articulation without affecting the spectral content: the pitch-contour and the time-evolution of the formant structure should be time-scaled, but otherwise not modified.

Time-scale warping function. Defining an arbitrary time-scale modification amounts to specifying a mapping between the time *in the original signal* and the time *in the modified signal*. This mapping $t \to t' = D(t)$ is referred to as *the time-scale warping function*; in the sequel t refers to the time in the original signal, t' to the time in the time-scaled signal. It is often convenient to give an integral definition of D:

$$t \to t' = D(t) = \int_0^t \beta(\tau)\, \mathrm{d}\tau, \tag{30}$$

in which $\beta(\tau) > 0$ is the time-varying *time-modification rate*. For a constant time-modification rate $\beta(\tau) = \beta$, the time-scale warping function is linear $t \to t = D(t) = \beta t$. The case $\beta > 1$ corresponds to slowing down the apparent rate of articulation by means of time-scale expansion, while $\beta < 1$ corresponds to speeding up the apparent rate of articulation by means of time-scale compression. For time-varying time-modification rates, the function $t \to t' = D(t)$ is non-linear. Note that $\beta(t)$ is implicitely assumed to be a regular and "slowly" varying function of time, i.e., its bandwidth is several orders of magnitude smaller than the effective bandwidth of the signal to be modified.

Ideal time-scale modification. The time-scale warping function specifies that "*speech events*" that occurred at time t in the original signal should take place at time $t' = D(t)$ in the time-scaled signal. With reference to the voiced speech production model introduced above, the speech parameters should be transformed the following way: [4]

$$n' \to P'(n') = P\big(D^{-1}(n')\big), \tag{31}$$

$$\begin{aligned} n' \to A'_k(n') &= G'_k(n') \\ &= G\big(D^{-1}(n'), \omega_k(D^{-1}(n'))\big), \\ 0 \leqslant k &\leqslant P'(n') - 1, \end{aligned} \tag{32}$$

$$\begin{aligned} n' \to \theta'_k(n') &= \Phi'_k(t') + \psi\left(D^{-1}(n'), \frac{2\pi k}{P(D^{-1}(n'))}\right), \\ n' \to \Phi'_k(n') &= \sum_{m=0}^{n'} \frac{2\pi k}{P(D^{-1}(m))}, \end{aligned}$$

where $D^{-1}(\cdot)$ denotes the inverse mapping (linking the new time-scale back to the original time-scale). These equations express that

[4] It is implicitely assumed that the time-modification rate factor is moderate, say $0.25 \leqslant \beta(t) \leqslant 4$.

1. the pitch contour $n' \rightarrow P'(n')$ of the modified signal is time-scaled by the warping function D;
2. the system function (amplitude $G'(n', \omega)$ and phase $\psi'(n', \omega)$) is a time-scaled version of the original system function;
3. the instantaneous frequency of the k-th harmonic in the modified signal at time n' corresponds to the instantaneous frequency in the original signal at time $D^{-1}(n')$; this is verified by evaluating the first-order backward difference of the instantaneous phase $\theta'_k(n')$.

2.3.2. Pitch-scale modification

The object of pitch-scale modifications is to alter the fundamental frequency of a speech segment without affecting its spectral enveloppe (more precisely, the locations and bandwidths of the formants) nor its time evolution.

Pitch warping function. Defining an arbitrary pitch-scale transformation amounts to specifying a time-varying pitch-modification factor $t \rightarrow \alpha(t) > 0$ which will affect the pitch-contour: the transformed pitch-contour $n \rightarrow P'(n)$ is

$$n \rightarrow P'(n) = \frac{P(n)}{\alpha(n)}. \tag{33}$$

When the pitch-transformation factor is larger than one, $\alpha(t) > 1$, the local pitch-frequency is increased by a factor $\alpha(t)$ (the local pitch-period is multiplied by a factor $1/\alpha(t)$); similarly, $\alpha(t) < 1$ indicates that the pitch-frequency is lowered by a factor $\alpha(t)$ (the pitch-period is lengthened accordingly). As above, $\alpha(t)$ is implicitely assumed to be a regular and "slowly" varying function of time.

Ideal pitch-scale modification. With reference to the speech production model, the speech parameters must be modified according to

$$n' \rightarrow P'(n') = \alpha(n')P(n'), \tag{34}$$

$$n' \rightarrow A'_k(n') = G'_k(n') = G(n', \alpha(n')\omega_k(n')), \tag{35}$$

$$n' \rightarrow \theta'_k(n') = \Phi'_k(n') + \psi(n', \alpha(n')\omega_k(n')),$$

$$n'\Phi'_k(n') = \sum_{m=0}^{n'} \alpha(m)\omega_k(m).$$

These equations can be interpreted as follows:

1. The pitch-contour is scaled by the time-varying factor $\alpha(n')$.
2. The amplitudes of the pitch-modified harmonics are obtained by sampling the system magnitude function at the shifted pitch-harmonic frequencies, thus preserving the formant structure.
3. The first-order backward difference of the modified excitation instantaneous phase Φ'_k is equal to the scaled k-th pitch-harmonic frequency: the pitch is indeed modified.

As opposed to time-scale modifications, pitch-scale modifications require the estimation of the system amplitudes $G(n', \alpha(n')\omega_k(n'))$ and phases $\psi(n', \alpha(n')\omega_k(n'))$, at frequencies not necessarily corresponding to a pitch-harmonic in the original signal. This is why most pitch-scale modification algorithms require the explicit decomposition of the speech signal into (i) a time-varying spectral envelope and (ii) a source (or excitation) signal with a flat spectrum. Because the time-varying system function $G(n,\omega)$ is *not exactly identifiable* from the input signal waveform, additional assumptions are needed for this task. The following section briefly describes several methods that can be used for this purpose.

2.4. Source-filter decomposition

Source-filter decomposition is a standard procedure in speech processing applications, including low bit-rate speech coding, speech recognition, or speech synthesis. Obviously this kind of decomposition may be achieved in many different ways, and a whole issue of this journal would not suffice to give a fair description of all the methods which have been proposed over the last 20 years. In this section, we only concentrate on two simple approaches, which are known to provide satisfactory results: [5] the first one is based on all-pole (LPC) modeling of speech signal, the second one makes explicit use of the short-time

[5] The pitch-scale modification algorithms described in this paper are not based on parametric models (e.g., glottal LPC vocoder) which are known to be sensitive to modeling errors. In the present context, limited envelope/source decomposition errors do not severely impair the quality of the processed speech.

amplitude spectrum: it is most easily implemented in procedures that explicitly use a frequency-domain representation of the signal.

2.4.1. All-pole (LPC) modeling

The LPC (Markel and Gray, 1976) method is perhaps the most widely used model for speech signal processing. In the LPC framework, the slowly-varying spectral envelope of the speech signal is estimated by computing the coefficients of an all-pole linear filter in short-time frames centered around the analysis time-instants (the typical frame size is 20–30 msec; 10 and 20 coefficients are used, depending on the speech bandwidth). Several estimation methods are documented in the litterature: the standard autocorrelation method (see for example (Markel and Gray, 1976)) is used most of the time because of its intrinsic simplicity (in particular, the stability of the synthesis filter need not be checked, as opposed to the covariance estimation procedure). Using this stream of LPC models, the source signal is determined by inverse filtering. To avoid problems in segments where the filter coefficients are changing rapidly, a standard trick consists of interpolating the filter coefficients, for example on a sample-by-sample basis. This interpolation may be performed either directly on the prediction coefficients (some care should be taken to avoid instability) or on "transformed" coefficients, such as the reflection or Log-Area Ratio coefficients. A normalized implementation of LPC filters can also be used. See for example (Markel and Gray, 1976; Depalle, 1991) for more detail.

Problems arise with standard LPC when processing high-pitched female voices because the LPC filter tends to model the individual pitch-harmonics rather than the overall envelope. In this case, alternate parameter estimation methods can be used, for example the "discrete all-pole modeling" originally proposed by ElJaroudi and Makhoul, which aims at fitting the LPC envelope directly on the pitch-harmonics, while imposing additional smoothness constraints to avoid over-modeling. Such techniques are described in (ElJaroudi and Makhoul, 1986, ElJaroudi, 1991; Depalle, 1991; Galas and Rodet, 1991).

2.4.2. Direct modeling

For frequency-domain implementations, direct modeling of the speech spectrum envelope may be considered, since the short-time analysis spectrum is used as an intermediate representation of the signal. The most straightforward approach in this case consists of interpolating the pitch-harmonics magnitudes. The interpolation may be linear or quadratic, but always requires either a fractional-delay pitch-period estimate or a preliminary peak picking procedure (McAulay and Quatieri, 1986). This envelope being real, it is by default zero-phase (the associated impulse response is symmetric about the time-origin). If necessary, the minimum-phase envelope may be calculated by use of the Hilbert-transform relations (Oppenheim and Schafer, 1989). In either case, the source short-term spectrum is obtained by dividing the short-term analysis spectrum by the spectral envelope. Special care should be taken to avoid time-aliasing, for example by increasing the number N of FFT points.

An interesting alternate solution consists of estimating the coefficients of the LPC filter whose transfer function best approximates the zero-phase envelope. This can be done by applying an inverse Fourier transform to the zero-phase squared envelope, then using the Levinson algorithm on the resulting pseudo autocorrelation sequence. Such procedures are described in (Depalle, 1991).

3. Time / pitch-scaling using the short-time Fourier transform

3.1. Time-scaling using the short-time Fourier transform

In this section, we will assume that the speech production model described in Section 2.2 is valid, and that the time-scale modification is specified by a time-scale warping function $t \rightarrow D(t)$, as in Section 2.3. The speech signal is analyzed using the short-time Fourier transform: it is further assumed that the analysis window and the FFT size are chosen so that the narrow-band analysis conditions are met, i.e. Eqs. (23), (24) and (25) hold.

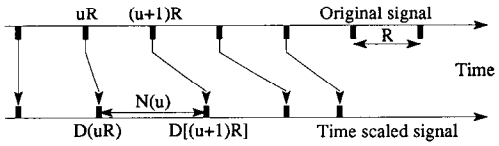

Fig. 2. Correspondance between the analysis time-instants uR and the synthesis instants $\tilde{t}_s(u) = D(t_a(u))$.

3.1.1. Time-scaled short-time Fourier transform Voiced speech segments. Although constant rate time-scale modifications can be performed using the same analysis and synthesis rates (Portnoff, 1981b), the algorithm is significantly simplified if the analysis is performed at a constant rate R different from the possibly time-varying synthesis rate.

To each analysis time-instant $t_a(u) = uR$ is associated a time-scaled virtual synthesis instant $\tilde{t}_s(u) = D(uR)$. Fig. 2 gives an illustration of the correspondance between fixed-rate analysis time-instants uR and "elastic" virtual synthesis time-instants $\tilde{t}_s(u) = D(t_a(u))$.

Now the virtual synthesis time-instants are not necessarily integers, depending on the analysis time-instants $t_a(u)$ and the time-scale warping function $D(t)$. However, the short-time Fourier synthesis scheme requires the values of $Y(t_s(u),\Omega_k)$ be given at times multiple of the sample period. To remedy this problem, the synthesis instants are shifted to the nearest integers below, and $Y(t_s(u),\Omega_k)$ calculated at time-instants $t_s(u) = \lfloor \tilde{t}_s(u) \rfloor$ in which $\lfloor x \rfloor$ designates the largest integer smaller than x.

The time-scaled short-time Fourier transform $Y(t_s(u),\Omega_k)$ at time-instants $t_s(u)$ is then given by the following equations [6]:

$$Y(t_s(u),\Omega_k) = M\left[D^{-1}(\tilde{t}_s(u)),\Omega_k\right] \times \exp\left[j\hat{\phi}(t_s(u),\Omega_k)\right],$$

with

$$\hat{\phi}(m,\Omega_k) = \hat{\phi}(t_s(u-1),\Omega_k) + \sum_{i=1}^{m-t_s(u-1)} \lambda\left((u-1)R + \frac{iR}{N(u)},\Omega_k\right)$$

$$\text{for } t_s(u-1) \leqslant m \leqslant t_s(u), \qquad (36)$$

$$N(u) = t_s(u) - t_s(u-1),$$

in which $N(u)$ is the time in samples between two successive synthesis time-instants, and $\lambda(n,\Omega_k)$ is the instantaneous frequency in the k-th channel at time n estimated according to Eq. (29). It is easily verified that (i) the instantaneous amplitude of the sinusoid in the k-th channel at time $t_s(u)$ in the time-scaled signal is equal to the corresponding instantaneous amplitude in the original signal at time $t_a(u)$, and (ii) the instantaneous frequency of the pitch-harmonic at time $t_s(u)$ in the time-scaled signal is

$$\hat{\phi}(t_s(u),\Omega_k) - \hat{\phi}(t_s(u-1),\Omega_k) = \lambda(uR,\Omega_k),$$

i.e., it is equal to the instantaneous frequency of the pitch-harmonic at time $t_a(u) = uR$ in the original signal. One notices that the instantaneous frequency $\lambda(m,\Omega_k)$ needs to be estimated at every instants $(u-1)R + iR/N(u)$. The amplitude, however, does not require any interpolation. In practice, provided the analysis time-instants are close enough, the instantaneous frequency of the pitch-harmonic in the k-th channel can be considered constant between $(u-1)R$ and uR. Based on this assumption, Eq. (36) simplifies to

$$\hat{\phi}t_s(u),\Omega_k) = \hat{\phi}(t_s(u-1),\Omega_k) + N(u)\lambda((u-1)R,\Omega_k).$$

The preceding equations can be interpreted as follows: the time-scaled short-time Fourier transform $Y(t_s(u),\Omega_k)$ is obtained by time-scaling the evolutions of the amplitudes and instantaneous frequencies, and letting the instantaneous phase run free as shown by Eq. (36). This guarantees that the time-scaled signal contains the same pitch-harmonics frequencies, and that their evolutions are time-scaled according to the desired time-scale warping function $t \to D(t)$.

[6] In his original paper (Portnoff, 1981b), Portnoff proposed a more sophisticated scheme in which the phase of the STFT was decomposed into the slowly varying system phase, and the excitation phase. It seems, however, that the simpler scheme presented here is the one used mostly in practice.

Finally, the time-scaled signal is obtained by feeding the time-scaled STFT $Y(t_s(u),\Omega_k)$ into the synthesis formula Eq. (9). Note that when the time-scale warping function $t \to D(t)$ is not linear (i.e., when the scale factor is not constant), the synthesis window must be normalized at each synthesis frame according to Eq. (9). In practice, however, provided the synthesis windows have sufficient overlap, this normalizing stage can be skipped.

Unvoiced speech segments. Although the modification system described above relies on the assumption that the speech signal is quasi-periodic with slowly time-varying parameters, the same system can also be applied as is to unvoiced signals. Using the same modification scheme for voiced or unvoiced speech segments adds to the robustness of the algorithm because no voiced/unvoiced decision is needed. However, applying the same method to both voiced and unvoiced segments generates artifacts for relatively large modification factors (above about 2): the time-scaled unvoiced speech segments acquire a "tonal" quality, a result of the phase coherence between successive synthesis short-time spectra (Eq. (36)).

3.1.2. Implementation

The algorithm can be summarized as follows:

1. Set the initial instantaneous phases $\hat{\phi}(0,\Omega_k) = \arg(X(0,\Omega_k))$.
2. Compute the short-time Fourier transform at next analysis time-instant $t_a(u)$ and calculate the instantaneous frequency in each channel according to Eq. (29).
3. Calculate next time-scaled synthesis instant $t_s(u) = \lfloor D(t_a(u)) \rfloor$ and the corresponding instantaneous phase $\hat{\phi}(t_s(u),\Omega_k)$ by

$$\hat{\phi}(t_s(u),\Omega_k) = \hat{\phi}(t_s(u-1),\Omega_k) + N(u)\omega_s((u-1)R),$$

with

$$N(u) = t_s(u) - t_s(u-1).$$

4. Reconstruct the time-scaled short-time Fourier transform at time $t_s(u)$ by

$$Y(t_s(u),\Omega_k) = A_k(u)\exp\left[j\hat{\phi}(t_s(u),\Omega_k)\right].$$

5. Calculate the u-th short-time modified signal by use of the synthesis formula Eq. (9) and return to step 2.

3.2. Pitch-scaling using the short-time Fourier transform

Pitch-scale modification using the phase-vocoder involves four steps; first, the source signal and the spectral envelope are extracted from the signal. Second, the source signal is processed to modify its pitch-structure, using a resampling procedure. Third, a pitch-modified speech signal is obtained by recombining the spectral envelope and the pitch-modified excitation. The resampling stage alters the duration of the segment: compensatory time-scale modification is applied in the fourth, final step to restore the original segment duration.

Source/filter decomposition is documented in Section 2.4. The other processing stages are described below. Because resampling is central to the pitch-scaling algorithms, it is presented in detail in the following section.

3.2.1. Time-domain and frequency-domain resampling

Two resampling methods are presented here, the first one working in the time-domain, the second one in the frequency domain. The latter can easily be incorporated in the phase-vocoder analysis/synthesis scheme.

Time-domain resampling method. The standard time-domain resampling method described here is generally used for *constant* and *rational* sampling-rate conversion factors $\alpha = D/U$. This resampling method consists of

1. upsampling the original signal by a factor U by inserting $U-1$ zero-valued samples between adjacent samples;
2. interpolating the resulting upsampled signal by use of a low-pass filter with an appropriate cutoff-frequency (this point is detailed below); and
3. downsampling the resulting signal by a factor D by discarding $D-1$ samples out of D.

The spectrum of the upsampled signal (obtained at the end of the first step above) is a U-folded

compressed version of the original spectrum: it is made of U copies (*images*) of the original spectrum compressed in the interval $[-\pi/U, \pi/U]$ (Crochiere and Rabiner, 1983).

Alias-free reconstruction of the upsampled signal requires the exact cancellation of these images, an operation performed during step 2 above. The interpolating filter must have a cutoff frequency [7] of π/U. However, the last step also requires low-pass filtering, with a cutoff frequency of π/D; the two filtering operations are usually performed at one time, with a cutoff frequency of $\min(\pi/D,\pi/U)$. Downsampling by a factor D, the ideally interpolated signal produces a decimated signal whose frequency contents is concentrated in the interval $[-\pi D/U, \pi D/U]$ when $D/U < 1$ and $[-\pi,\pi]$ when $D/U > 1$.

The time-domain source resampling method is straightforward to implement when one is working with (i) constant and (ii) rational sampling-rate conversion factors: a single (or a simple cascade of) low-pass interpolating filter is all that is required. The situation is quite different when the sampling-rate conversion factor is (i) time-varying or (ii) irrational. The phase vocoder provides an elegant solution to this problem, operating as a "non-canonical", frequency-domain sampling-rate converter, as described in the following section.

Frequency-domain resampling method. Sampling rate conversion by arbitrary (possibly time-varying) ratios $\alpha(t) \geqslant 0$ is an important application of STFT analysis–synthesis. In the STFT analysis–synthesis framework, the modification of the sampling period is carried out in the frequency-domain and involves the following three steps:

1. The extraction of a stream of short-term analysis signals at a uniform analysis rate of R samples, i.e. $t_a(s) = sR$ and the computation of the associated short-term spectra.
2. The interpolation of the complex short-term Fourier spectra at a new set of frequencies, and the computation of the associated short-term synthesis signals.
3. The synthesis of the sampling-rate modified signal.

During step 2 above, linear interpolation is applied to the real and the imaginary parts of the short-term spectrum. The linearly interpolated short-term spectrum is expressed as

$$\begin{aligned}\bar{X}(\tilde{t}_s(u),\Omega_k) &= (1-\rho(k))\,X(t_a(u),\Omega_{\tilde{k}}) \\ &\quad + \rho(k)\,X(t_a(u),\Omega_{\tilde{k}+1}), \\ \tilde{k} &= \lfloor k\alpha(t_a(u)) \rfloor, \\ \rho(k) &= k\alpha(t_a(u)) - \tilde{k},\end{aligned} \tag{37}$$

where $\lfloor x \rfloor$ indicates the largest integer smaller than x.

Resampling the signal by a factor $\alpha(t)$ (be it in the time or in the frequency domain) causes a local modification of the time-scale by a factor $\beta(t) = 1/\alpha(t)$. As a consequence, the short-time modified signal corresponding to $\bar{X}(\tilde{t}_s(u),\Omega_k)$ is now synchronized on the virtual synthesis time-instant $\tilde{t}_s(u)$ given by

$$\tilde{t}_s(u) = \int_0^{uR} 1/\alpha(t)\,\mathrm{d}t, \tag{38}$$

and the effective length of the short-time modified signal is divided by $\alpha(uR)$. Note that when increasing the sampling rate, the duration of the short-time modified signal is *increased* and can exceed the length N of the Fourier transform, causing time-alisasing in the resampled output. To avoid this problem, the length T_u of the analysis window $h_u(n)$ should always be smaller than $N\alpha(t_a(u))$. No such problem is encountered, when lowering the sampling rate.

Because of the generally non-integral nature of sample rate change, fractional delays occur between synthesis frames. More precisely, the virtual synthesis time-instants $\tilde{t}_s(u) = D(uR)$ do not necessarily correspond to integer numbers of samples, although the synthesis scheme requires integer synthesis time-instants $t_s(u) = \lfloor \tilde{t}_s(u) \rfloor$. In order to correct the *fractional delay* $\tilde{t}_s(u) - t_s(u)$, it is necessary to apply a fractional-sample delay to each frame so that frames overlap with proper phase relation. These delay adjustments are ap-

[7] The interpolation filter design may be found in standard signal processing textbooks, see for example (Lim and Oppenheim, 1988).

plied on each short-term spectrum as a phase correction:

$$\hat{Y}(t_s(u),\Omega_k) = \overline{X}(\tilde{t}_s(u),\Omega_k) \times \exp\left[-i\Omega_k(\tilde{t}_s(u) - t_s(u))\right]. \quad (39)$$

The resampled output is finally obtained by a weighted least-square overlap-add procedure (see Eqs. (11) and (9)); note that, due to the time-varying nature of the modification, the synthesis rate is not constant whereas the analysis rate is.

During the resampling process, the cutoff-frequency ω_h of the analysis window is multiplied by a factor $\alpha(t_a(u))$. To avoid artifacts during the synthesis, the bandwidth of the synthesis window should be matched to the modified cutoff-frequency $\alpha(t_a(u))\omega_h$. For standard spectral analysis windows (e.g., Hanning, Hamming, Kaiser), the synthesis window is

$$f_u(n) = h_u(n/\alpha(t_a(u))).$$

For simple decimation factors (ratios of small integers, e.g., 2/3 or 4/5), the technique described here is somewhat more computationally involved than the standard time-domain interpolation/decimation procedure described above. This explains its limited use in practice. However, this method is quite flexible and allows time-varying modification rates, an advantage which is fully exploited for pitch-scale modification.

3.2.2. *Pitch-scale modification based on resampling*

In the context of pitch modification, resampling is performed *without modifying the actual sampling frequency*: the normalized frequency interval $[-\pi, +\pi]$ remains associated with the same "physical frequencies". As a consequence upsampling and downsampling operations result:

- in linear compression-expansion of the frequency axis,
- in linear expansion-compression of the time axis.

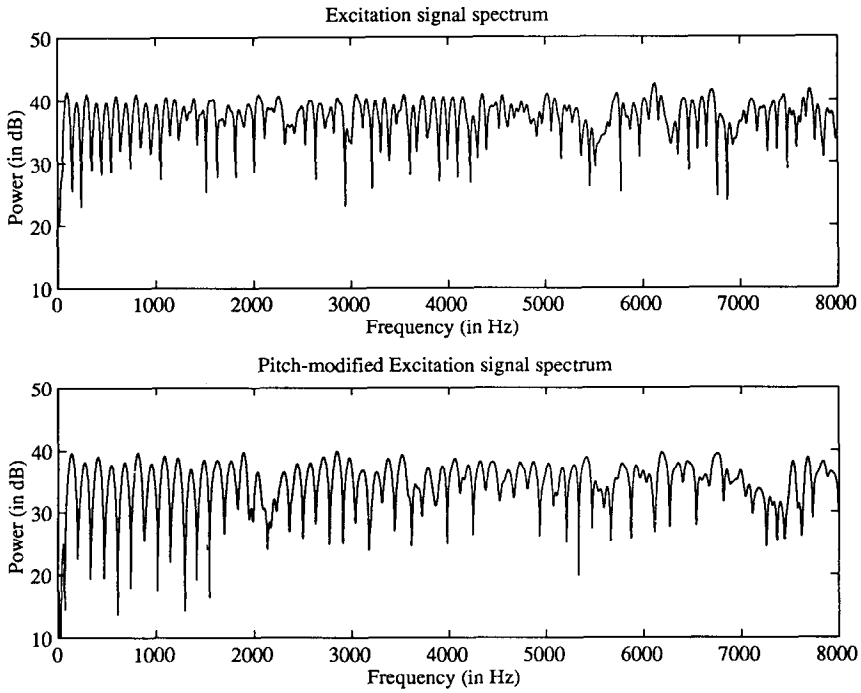

Fig. 3. Pitch-modification by time-domain source signal resampling. Upper panel: short-time amplitude spectrum of the original source signal; lower panel: short-time amplitude spectrum of the resampled source signal. The pitch-scale factor is constant and equal to 4/3. Note that the source spectrum is a simple expansion of the original spectrum.

When applied to speech signals, the compression-expansion of the frequency axis modifies the spacing between successive pitch-harmonics (pitch-modification, see Fig. 3) but also modifies the locations and bandwidths of the formants. This, of course, is undesirable for speech pitch-scale modification. To circumvent this problem, resampling is carried out on the source signal obtained by a source/filter decomposition method (see Section 2.4).

The linear expansion-compression of the time axis must also be compensated by a time-scale modification technique. Finally, the desired pitch-modified waveform is obtained by combining the resampled source signal and the unprocessed envelopes.

High-frequency regeneration. When one increases the pitch of the source ($\alpha = D/U > 1$), the spectrum of the source signal is expanded and the upper frequency part of the source spectrum is simply discarded.

By contrast, as mentioned in Section 3.2.1, when one decreases the pitch of the source signal ($\alpha < 1$), its spectrum only occupies a limited portion ($[-\pi\alpha, \pi\alpha]$) of the frequency range. If this source signal is used as is in the synthesis stage, the resulting pitch-scaled speech signal is artificially band-limited. To counter this undesirable side-effect, the missing upper-frequency band of the source signal needs to be regenerated before the final synthesis stage.

Two simple methods have been proposed for high-frequency regeneration: *spectral folding* and *spectral copying*.

Spectral folding is a time-domain method which basically consists of using an *incorrect* cut-off frequency at the interpolation stage (step 2 in Section 3.2.1) to allow a certain amount of frequency aliasing.

Assume that the interpolating low-pass filter has a cutoff frequency of π/D. The cutoff frequency being larger than the limit π/U guaranteeing the exact cancellation of the images, the resulting signal contains some frequency aliasing. Decimating the signal by a factor D yields a signal occupying the whole normalized frequency interval $[-\pi, +\pi]$: the upper frequencies are folded versions of the compressed original spectrum.

The main advantage of the spectral folding method is obviously its simplicity: it requires *no additional computation* to regenerate the high frequency components.

Spectral copying can be performed only with frequency-domain resampling methods. The technique involves copying the lower-frequency components of the spectrum into the upper part. The duplication of the lower part of the spectrum was first suggested by Seneff (1982). Duplication is somewhat more involved than spectral folding, since the phase of each individual frequency component needs to be adjusted. This is especially true when dealing with reduced-bandwidth signals (8 kHz sampling-rate); it is much less critical for wide-band speech (16 kHz and above) since the missing frequencies, corresponding to turbulent airflow friction noise, are most likely unvoiced (Griffin and Lim, 1984). Details can be found in (Seneff, 1982).

Compensatory time-scaling. In order to give back its original time-evolution to the pitch modified signal, a compensatory time-scale modification must be performed that maps the virtual synthesis time-instants $\tilde{t}_s(u)$ back to the original analysis time-instants $t'_s(u) = uR$. According to Eq. (38) the corresponding time-scale warping function is $D^{-1}(t)$, where $D(t)$ is given by

$$D(t) = \int_0^t 1/\alpha(u)\,\mathrm{d}u. \tag{40}$$

3.2.3. Phase-vocoder implementation of pitch-scale modification

The phase-vocoder offers a natural framework for pitch-scale modification by frequency domain resampling because it simplifies significantly the successive steps of the modification. In particular, the source resampling and the compensatory time-scale modification can be performed at one time, with no additional computation. More precisely, rather than synthesizing the pitch-modified source signal, then applying compensatory time-scaling, it is more efficient to incorporate the time-scaling into the STFT synthesis stage.

It is easy to verify that the required time-scale modification expressed by Eq. (40) corresponds to a time-varying analysis rate and a constant syn-

thesis rate: the virtual, non-uniform analysis marks $\tilde{t}_s(u)$ must be mapped to the uniform synthesis marks $t'_s(u) = uR$.

The phase-vocoder method for time-scale modification presented in Section 3.1 can easily be adapted to the present context.

The phase-vocoder implementation of the source-resampling pitch-scaling method can then be summarized as follows:

1. Using any of the methods described in Section 2.4, perform a source-filter decomposition of the original signal.
2. Perform a short-time Fourier transform on the source signal. Note that the two first steps can be carried out simultaneously, for example when direct modeling is used. At this point, the original signal is decomposed into the short-time spectra of the excitation $X_s(uR,\Omega_k)$ and a stream of time-varying envelope functions $G(uR,\omega)$.
3. Calculate the linearly resampled source-spectrum $\bar{X}_s(\tilde{t}_s(u),\Omega_k)$ by use of Eq. (37).
4. If needed (i.e. if $\alpha(t_a(u)) < 1$), perform spectral copying or spectral folding to regenerate the high-frequency components.
5. In each FFT channel, express $\bar{X}_s(\tilde{t}_s(u),\Omega_k)$ in polar coordinates,

$$\bar{X}_s(\tilde{t}_s(u),\Omega_k) = \bar{M}(\tilde{t}_s(u),\Omega_k) \times \exp\left(j\bar{\phi}(\tilde{t}_s(u),\Omega_k)\right),$$

and calculate the instantaneous frequency $\lambda(\tilde{t}_s(u),\Omega_k)$ using $\bar{\phi}(\tilde{t}_s(u),\Omega_k)$ and $\bar{\phi}(\tilde{t}_s(u-1),\Omega_k)$ as seen in Section 2.2.2. Note that $R(u) = t_a(u) - t_a(u-1)$ must be replaced by $\tilde{t}_s(u) - \tilde{t}_s(u-1)$.

6. Calculate the pitch-scaled instantaneous phase by

$$\bar{\phi}'(uR,\Omega_k) = \bar{\phi}'((u-1)R,\Omega_k) + R\lambda(\tilde{t}_s(u),\Omega_k).$$

7. Reconstruct the pitch-scaled source spectra $Y_s(uR,\Omega_k)$ by

$$Y_s(uR,\Omega_k) = \bar{M}(\tilde{t}_s(u),\Omega_k)\exp\left[j\bar{\phi}'(uR,\Omega_k)\right].$$

8. Calculate the pitch-scaled source signal $y_s(n)$ by use of the synthesis formula Eq. (9) with $t_s(u) = uR$.
9. Apply the time-varying filters $G(uR,\omega)$ to the pitch-scaled source signal. The output is the desired pitch-scaled signal.

The two last steps can be swapped, for example when direct modeling is used.

A popular, perhaps simpler way of resynthesizing the pitch-modified signal consists of replacing the overlap-add synthesis stage by a parallel bank of sinusoidal oscillators. Having performed a short-time Fourier transform on the original speech signal, the instantaneous frequencies $\lambda(\tilde{t}_s(u),\Omega_k)$ are calculated in each channel, then multiplied by the desired pitch-modification factor. The corresponding sinusoidal amplitudes $A_k(uR)$ are derived from the envelope function, i.e. $A_k(uR) = G(uR, \lambda(\tilde{t}_s(u),\Omega_k)) \times \bar{M}(\tilde{t}_s(u),\lambda(\tilde{t}_s(u),\Omega_k))$, and a bank of sinusoidal oscillators in parallel is used to generate the pitch-scaled output signal. The oscillators are controlled in frequency and amplitude, and their parameters ($A_k(uR)$ and $\lambda(\tilde{t}_s(u),\Omega_k)$) are updated at times uR. This scheme has the advantage of bypassing the resampling of the excitation short-time spectrum (step 3 above). However, the oscillator-bank synthesis stage is usually more time-consuming than overlap-add synthesis.

3.2.4. Potential drawbacks of pitch-scale modification by resampling

The resampling methods present several drawbacks, which are summarized below:

High-frequency regeneration. Be it by spectral folding or by spectral copying, the technique used to regenerate the upper part of the spectrum is not satisfactory. The problem is very serious when reduced band signal are processed (8 kHz sampling frequency). This point illustrates a primary motivation for substituting a more appropriate technique such as the pitch-synchronous TD-PSOLA method or parametric methods.

Harmonic amplitude deviations. Spectral envelope estimation methods are not "perfect" in the sense that they generally fail to flatten completely the spectrum of the source signal. The "flattened" pitch-harmonic amplitudes present certain fluctu-

ations with respect to the "ideal" flat envelope spectrum. [8] Time-domain or frequency-domain source resampling techniques linearly warp the frequency axis: the amplitude deviations of the pitch-harmonics are also shifted in frequency. *Voicing.* The other details of the source are also shifted from their original spectral location. This is particularly true for the voicing, which is not uniform over the frequencies. The source spectrum of a voiced sound usually exhibits frequency-bands where friction noise dominates the pitch-harmonics (this phenomenom is more pronounced in the upper frequency part of the spectrum). The pitch-modification methods based on frequency-domain resampling preserve the frequency-dependent voicing (which is obviously desirable), but translate the voiced/unvoiced frequency bands (which may degrade the quality).

4. The PSOLA system

Most time-domain techniques for time-scale and pitch-scale modification are based on the same fundamental method, best illustrated by the Pitch-Synchronous Overlap and Add method (PSOLA). We start by describing the main characteristics of the PSOLA technique: the decomposition of the input waveform into a stream of analysis short-time signals, the modification of each short-time analysis signal into a short-time synthesis signal, and the final overlap-add synthesis. We then document several variants of this basic scheme, including the less popular but equally effective frequency-domain version of the algorithm (referred to as the frequency-domain PSOLA, or FD-PSOLA for short) and different combinations of the basic PSOLA scheme with more traditional vocoder techniques (in particular, the linear-predictive PSOLA, or LP-PSOLA). A brief theoretical interpretation of the PSOLA method is finally presented.

4.1. The PSOLA analysis–synthesis framework

As outlined above, the overall PSOLA analysis-modification-synthesis scheme can be decomposed in three successive steps.

4.1.1. From speech waveform to short-time analysis signals

The first step of the analysis process consists of decomposing the speech waveform $x(n)$ into a stream of short-time analysis signals, denoted $x(s,n)$ (as previously, s is the index of the short-time signal and n is the sample index within the short-time signal). These short-time signals are obtained by multiplying the signal waveform $x(n)$ by a sequence of time-translated analysis windows (see Section 2.1): this is expressed as

$$x(s,n) = h_s(n)x(n - t_a(s)), \tag{41}$$

where $h_s(m)$ is the analysis window (whose support is located around the time-origin $n = 0$) and $t_a(s)$ is the s-th analysis time-instant. In the PSOLA context, the time-instants $t_a(s)$ are also referred to as the *analysis pitch-marks*; these pitch-marks are set at a pitch-synchronous rate on the voiced portions of speech and at a constant rate on the unvoiced portions.

The length of the analysis window T is proportional to the local pitch-period $P(s)$, i.e., $T = \mu P(s)$. The proportionality factor μ ranges from 2, for the standard time-domain PSOLA method, to 4, for the frequency-domain implementation. The analysis window $h_s(n)$ can be chosen arbitrarily: the key point is to use a window with a reasonable spectral behavior in terms of main-lobe bandwidth and side-lobe attenuation. Section 4.5 provides more information about the choice of $h(n)$ and μ and their influence on the quality of the output signal. In most implementations, $h_s(m)$ is a Hanning window.

4.1.2. From short-time analysis signals to short-time synthesis signals

The second step consists of transforming the stream of short-time analysis signals into a stream

[8] The actual magnitude of these deviations is difficult to evaluate, since it depends critically on the method used and on the number of model coefficients. For a standard LPC autocorrelation method with a reasonable number of coefficients (e.g. 16 for 16 kHz sampling frequency), these deviations may reach 5–10 dB, which is far from being negligible.

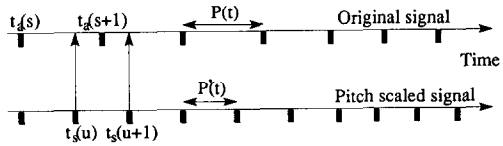

Fig. 4. Calculation of the synthesis pitch-marks for pitch-scale modifications. We have $P'(t) = P(t)/\beta$, with $\beta = 3/2$.

of short-time synthesis signals, synchronized on a new set of time instants $t_s(u)$, referred to as *the synthesis pitch-marks*.

The stream of synthesis pitch-marks $t_s(u)$ is determined from the analysis pitch-marks $t_a(s)$ according to the desired pitch-scale and time-scale modification. This point is crucial in the PSOLA method but has not yet been thoroughly documented in the literature; a precise description is given in Section 4.2. Along with the stream of synthesis pitch-marks, a mapping $t_s(u) \rightarrow t_a(s)$ between the synthesis and the analysis pitch-marks is determined, specifying which short-time analysis signal(s) should be *selected* for any given synthesis pitch-mark. Examples of such modifications and mappings are shown in Figs. 7 and 4 for constant time-scale and pitch-scale modification factors.

In the simplest implementation, a one-to-one correspondence between analysis and synthesis short-time signals is imposed: the stream of short-time synthesis signals is simply obtained by *eliminating or duplicating* analysis short-time signals. Assuming that the synthesis pitch-mark $t_s(u)$ is mapped with the analysis pitch-mark $t_a(s)$, the associated synthesis short-time signal $y(u,n)$ is simply $y(u,n) = x(s,n)$.

In more sophisticated systems, the mapping is no longer one-to-one but involves linear interpolation between the two successive short-time analysis signals lying the closest to the virtual pitch-mark associated with $t_s(u)$ (see below): the synthesis short-time signal $y(u,n)$ is then

$$y(u,n) = \alpha_u x(s,n) + (1-\alpha_u)x(s+1,n), \quad (42)$$

where $0 \leqslant \alpha_u \leqslant 1$ is a linear interpolation factor (the determination of the factor α_u is detailed below).

In the frequency-domain implementation (FD-PSOLA), the short-time synthesis signals are expressed in the frequency-domain, making this scheme computationally less attractive than the standard time-domain PSOLA approach which simply by-passes this modification stage. Frequency-domain modifications closely parallel those used in the conventional phase vocoder approach: these techniques are detailed in Section 4.6.

4.1.3. From short-time synthesis signals to the final output

In the final step, the synthetic signal $y(n)$ is obtained by combining the synthesis waveforms synchronized on the stream of synthesis pitch-marks $t_s(u)$. Weighted least-square overlap-add procedures may be used for this purpose (see Eqs. (9), (11) and (12)). In the time-domain PSOLA algorithm, the synthesis window $f_u(n)$ is equal to the analysis window associated with the analysis pitch-mark $t_a(s)$ mapped with the synthesis pitch-mark $t_s(u)$. A different synthesis window must be used in the frequency-domain PSOLA to take into account the inherent change in the time-scale introduced by the modification of the frequency axis (this is explained in Section 4.6.).

4.2. Computation of synthesis pitch-marks

The computation of the synthesis pitch-marks is done in two successive steps. First, the pitch-marks are generated according to the desired pitch-scale and time-scale modification, then each synthesis pitch-mark is associated with one or several analysis pitch-marks.

4.2.1. Pitch-scale modification

For pitch-scale modification, the stream of synthesis pitch-marks $t_s(u)$ is computed from the stream of the analysis pitch-marks $t_a(s)$ and the pitch-scale modification factors $\beta_s = \beta(t_a(s))$ in the following way. Assuming for simplicity that the signal is entirely voiced, the analysis pitch-marks $t_a(s)$ are positioned in a pitch-synchronous way, i.e. $t_a(s+1) - t_a(s) = P(t_a(s))$, in which $P(t)$ is a piecewise constant pitch contour function $t \rightarrow P(t)$,

$$P(t) = P(t_a(s)), \quad t_a(s) \leqslant t < t_a(s+1). \quad (43)$$

The synthesis pitch-marks must also be positioned pitch-synchronously, with respect to the synthesis pitch contour $t \to P'(t)$. We are left with the problem of finding a series of synthesis pitch marks $t_s(u)$ such that $t_s(u+1) = t_s(u) + P'(t_s(u))$ and $P'(t_s(u))$ is approximately equal to $1/\beta(t_s(u))$ times the pitch in the original signal around time $t_s(u)$:

$$P(t_s(u)) \approx \frac{P(t_s(u))}{\beta(t_s(u))}.$$

This is easily done recursively: we seek the value of $t_s(u+1)$ that satisfies

$$t_s(u+1) - t_s(u) = \frac{1}{t_s(u+1) - t_s(u)} \int_{t_s(u)}^{t_s(u+1)} \frac{P(t)}{\beta(t)}\, dt, \qquad (44)$$

$$\beta(t) = \beta(t_a(s)) = \beta_s \quad \text{for } t_a(s) \leqslant t < t_a(s+1). \qquad (45)$$

According to this equation, the synthesis pitch period $t_s(u+1) - t_s(u)$ is equal to the mean $1/\beta(t)$-scaled pitch period in the original signal calculated over the time-frame $t_s(u+1) - t_s(u)$. This integral equation in $t_s(u+1)$ is easily solved because $P(t)$ and $\beta(t)$ are piecewise constant functions. Fig. 4 illustrates the calculation of the synthesis pitch marks for pitch-scale modifications.

Pitch-scale modification using TD-PSOLA is shown in Figs. 5 and 6.

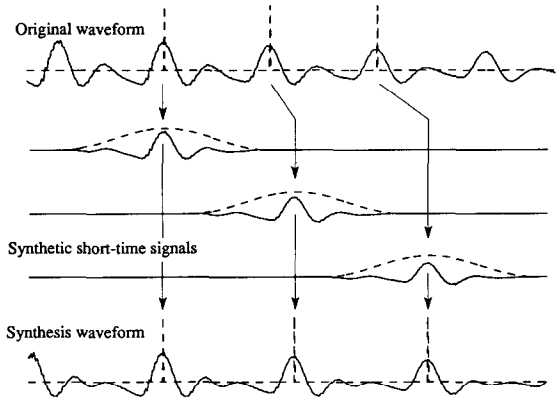

Fig. 5. Pitch-scale modification with the TD-PSOLA method. Upper panel: original signal along with the analysis pitch-marks. Middle panel: three short-time synthetic signals. The mapping between these short-time signals and the analysis pitch-marks are indicated by the arrows. Lower panel: pitch-scale modified waveform along with the synthetic pitch-marks. The signal is a vowel /i/, uttered by a male speaker (pitch frequency around 100 Hz). The pitch-scale modification factor is equal to 0.8.

4.2.2. Time-scale modification

Time-modifications are slightly more complicated than pitch-scale modifications because the original signal and the scaled signal do not share the same time-axis.

The time-scale modification is specified by associating to each analysis pitch-mark, a time-scale modification factor denoted $\alpha_s > 0$, from which the time-scale warping function $t \to D(t)$ may be deduced:

$$D(t_a(1)) = 0,$$
$$D(t) = D(t_a(s)) + \alpha_s(t - t_a(s)), \quad t_a(s) \leqslant t < t_a(s+1). \qquad (46)$$

$t \to D(t)$ is a piecewise linear and strictly monotonic function. Having specified the time-scale warping function, the next step consists of generating a stream of synthesis pitch-marks $t_s(u)$ from the stream of the analysis pitch-marks $t_a(s)$ *while preserving the pitch contour*. As in the preceding case, the analysis pitch-marks $t_a(s)$ are positioned in a pitch-synchronous way, i.e. $t_a(s+1) - t_a(s) = P(t_a(s))$. The target synthesis pitch-contour is defined as $t \to P'(t) = P(D^{-1}(t))$: the pitch in the time-scaled signal at time t should be close to the pitch in the original signal at time $D^{-1}(t)$.

We must now find a stream of synthesis pitch marks $t_s(u)$, such that $t_s(u+1) = t_s(u) + P'(t_s(u))$. To solve this problem, it is useful to define a stream of *virtual pitch-marks* $t'_s(u)$ in the original signal related to the synthesis pitch-marks by

$$t_s(u) = D(t'_s(u)), \qquad t'_s(u) = D^{-1}(t_s(u)).$$

Assuming that $t_s(u)$ and $t'_s(u)$ are known, we try to determine $t_s(u+1)$ (and $t'_s(u+1)$), such that $t_s(u+1) - t_s(u)$ is approximately equal to the pitch in the original signal at time $t'_s(u)$. This can be expressed as

$$t_s(u+1) - t_s(u) = \frac{1}{t'_s(u+1) - t'_s(u)} \int_{t'_s(u)}^{t'_s(u+1)} P(t)\, dt$$
$$\text{with } t_s(u+1) = D(t'_s(u+1)). \qquad (47)$$

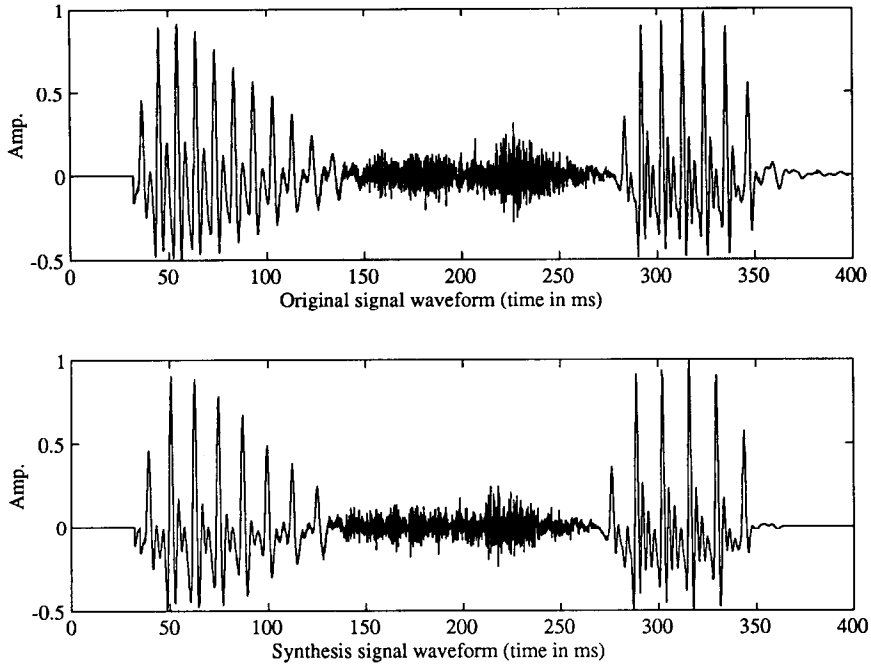

Fig. 6. Upper panel: original waveform. Lower panel: pitch-scale modified waveform. The pitch-scale modification factor is 0.8.

According to this equation, the synthesis pitch period $t_s(u+1) - t_s(u)$ at time $t_s(u)$ is equal to the mean value of the pitch in the original signal calculated over the time-interval $t'_s(u+1) - t'_s(u)$. Note that this time-interval $t'_s(u+1) - t'_s(u)$ is mapped to $t_s(u+1) - t_s(u)$ by the mapping function $D(t)$.

As was the case above, Eq. (47) is an integral equation but is easily solved because $D(t)$ and $P(t)$ are piecewise linear functions. Fig. 7 illustrates the calculation of the synthesis pitch marks for time-scale modifications. Time-scale modifi-

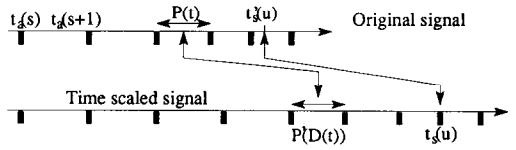

Fig. 7. Calculation of the synthesis pitch-marks for time-scale modifications. We have $P'(D(t)) \approx P(t)$ and $t_s(u) = D(t'_s(u))$.

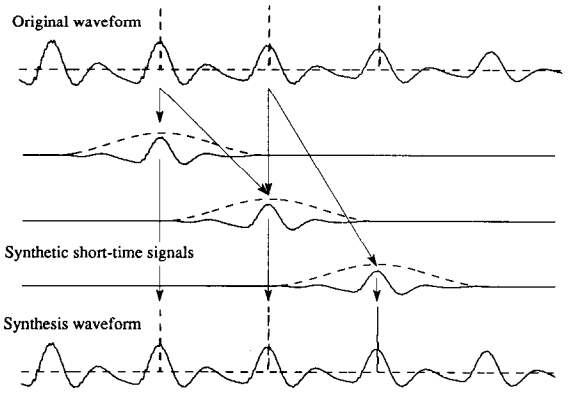

Fig. 8. Time-scale modification with the TD-PSOLA method. Upper panel: original signal along with the analysis pitch-marks. Middle panel: three short-time analysis signals. The mapping between these analysis signals and the associated analysis pitch-marks are indicated by the arrows. Lower panel: time-scale modified waveform along with the synthetic pitch-marks. Same signal as in Fig. 5. The time-scale modification factor is set to 2; the interpolation mode is used: the intermediate short-time signal is obtained by linearly interpolating the two adjacent waveforms, with a coefficient $\alpha = 0.5$.

cation using TD-PSOLA is shown in Figs. 8 and 9.

4.2.3. Combined time-scale and pitch-scale modifications

Combined time-scale and pitch-scale modifications follow easily from the preceding sections. Assuming as above that the pitch-scale modification and the time-scale modification are defined by two streams β_s and α_s, Eq. (47) can still be used after replacing $P(t)$ by $P(t)/\beta(t)$:

$$t_s(u+1) - t_s(u) = \frac{1}{t'_s(u+1) - t'_s(u)} \int_{t'_s(u)}^{t'_s(u+1)} \frac{P(t)}{\beta(t)}\, dt, \qquad (48)$$

$$\beta(t) = \beta_s \quad \text{with } t_a(s) \leqslant t < t_a(s+1). \qquad (49)$$

Combined time-scale and pitch-scale modifications actually involve no additional complexity. An example of such modifications is given in Fig. 10.

4.2.4. Mapping the synthesis pitch-marks to the analysis pitch-marks

The final step consists of associating a synthesis short-time signal to every synthesis pitch-mark (that is, working out a mapping between the synthesis and the analysis pitch-marks). It is clear from the previous discussion that a synthesis pitch-mark is not, in general, univoquely associated with an analysis pitch-mark: the virtual

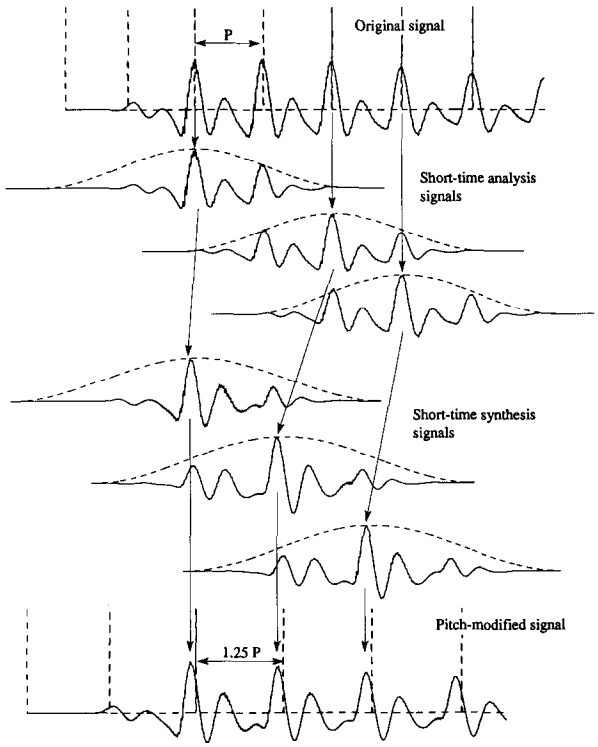

Fig. 10. Combined pitch-scale and time-scale modification using the frequency-domain PSOLA approach. The signal is the vowel /i/ uttered by a male speaker. The pitch-scale modification factor is 0.8 and the time-scale modification factor is 0.6; the spectral envelope is modeled by LPC (autocorrelation method) using 25 coefficients. The FFT size is 2048. Note that the synthesis time-instants no longer correspond to the maximum of the waveform.

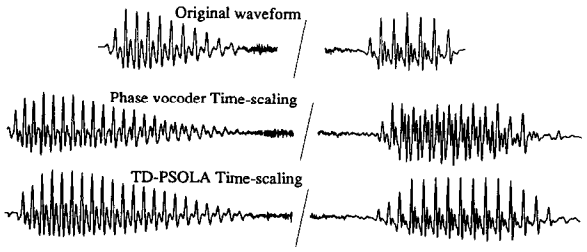

Fig. 9. Comparison between phase-vocoder time-scaling and TD-PSOLA time-scaling. Upper panel: original waveform. Middle panel: time-scaled signal obtained by the phase-vocoder technique. Lower panel: time-scaled signal obtained by the TD-PSOLA technique. The time-scale modification factor is 2.

pitch-mark $t'_s(u)$ (which is equal to the synthesis pitch-mark $t_s(u)$ for pitch-scale modifications) does not necessarily correspond to an analysis pitch-mark. A simple solution consists of taking a weighted average of the two closest analysis short-time signals. Suppose $t_a(s) \leqslant t'_s(u) < t_a(s+1)$, then

$$y(u,n) = (1-\alpha_u)x(s,n) + \alpha_u x(s+1,n),$$

$$\alpha_u = \frac{t'_s(u) - t_a(s)}{t_a(s+1) - t_a(s)}. \qquad (50)$$

In the simplest implementation, the coefficient α_u is replaced by the integer lying nearest to it,

that is 0 or 1. This latter solution is equivalent to selecting the analysis short-time signal associated with the analysis pitch-mark lying closest to the virtual pitch-mark $t'_s(u)$.

In that case, and when only time-scale modifications are undertaken, the time-domain and frequency-domain PSOLA methods work by eliminating or duplicating analysis short-time signals at a pitch-synchronous rate very much like most standard time-domain pitch-synchronous splicing methods (see Section 5).

4.3. Time-scaling of unvoiced sounds

In the preceding sections, we focused mainly on voiced speech modification. Special care should be taken when dealing with unvoiced speech segments. Recall that the PSOLA algorithm for time-scaling works by eliminating/duplicating short-time waveforms (with an optional smoothing of successive short-time signals). Unfortunately, when lengthening unvoiced fricatives, this process introduces a tonal noise because the repetition of segments of a "noise-like" signal produces an artificial long-time autocorrelation in the output signal, perceived as some sort of periodicity. A simple solution to this problem consists of reversing the time-axis whenever the algorithm needs to repeat a short-time signal, i.e. $x_s(m)$ is replaced by $x_s(-m)$. Such an operation preserves the short-time amplitude spectrum but changes the sign of the phase spectrum, thus reducing the undesirable correlation in the output signal. This approach efficiently eliminates tonal noise when the time-scale factor is less than 2 (this is usually enough for the modifications needed in practice). Precautions must also be taken with unvoiced plosives, in order to avoid degrading the transitional portions of the signal (bursts).

4.4. The frequency resampling property

Time-domain PSOLA is an astonishingly simple method for time-scale and pitch-scale modifications of speech signals.

The high quality of the obtained time-scaled signals is a consequence of the fact that the method operates by "smoothly" duplicating or eliminating parts of the speech signal at a pitch-synchronous rate. A complete theoretical analysis of the TD-PSOLA method is given in (Moulines and Charpentier, 1990). In this section, we describe the most significant results, the next section gives more insight into the choice of the analysis window.

For simplicity, we will assume that: (H1) The speech signal is purely periodic, with an integer pitch period P. (H2) The pitch-scale modification factor β is constant and is an integer number. (H3) The time-scale modification factor is constant and $\alpha = 1/\beta$. These hypotheses are idealistic, but the results can be extended to less stringent assumptions.

As a direct consequence of the hypotheses,

1. the analysis pitch-marks are given by $t_a(s) = sP$;
2. there is a one-to-one mapping between the analysis and the synthesis pitch-marks, i.e. $t_s(s) = s\beta P$;
3. the analysis windows (whose length is proportional to the local pitch period) are all equal to the same prototype window.

According to the preceding remarks, we have

$$x(s,n) = h(n)x(n - sP). \tag{51}$$

Assume for simplicity that we use the OLA procedure Eq. (11). Omitting the time-varying normalization factor, we get

$$y(n) = \sum_s x(s, n - s\beta P)$$

$$= \sum_s h(n - s\beta P)x(n - sP - s\beta P).$$

Since the input signal is assumed to be periodic with period P, $x(m - sP) = x(m)$, the preceding formula becomes

$$y(n) = \sum_s h(n - s\beta P)x(n - s\beta P). \tag{52}$$

The synthetic signal is thus obtained by replicating, with the period βP the *same* prototype signal, $h(n)x(n)$. Obviously, $y(n)$ is periodic with period βP. Standard results on the Fourier trans-

form of periodically extended waveforms yield the fundamental result:

$$y(n) = \frac{1}{\beta P} \sum_{k=0}^{\beta P-1} c_k \exp\left(\mathrm{j}\frac{2\pi k}{\beta P} n\right),$$

$$c_k = X\left(\frac{2\pi k}{\beta P}\right),$$

$$\text{with } X(\omega) = \sum_{n=-\infty}^{+\infty} h(n)x(n)\exp(-\mathrm{j}\omega n). \quad (53)$$

The above expression means that *the complex amplitudes c_k of the pitch-harmonics in the synthetic signal are equal to the values at the pitch-harmonic frequencies $2\pi k/\beta P$ of the discrete Fourier transform (DFT) of the prototype short-time signal $h(n)x(n)$*. In other words, the time-domain PSOLA method works by resampling the Fourier transform of the short-time analysis signal (this is illustrated in Figs. 11 and 12). This relation contains much useful information, and clarifies the influence of different settings (window length and window type) on the performance of the algorithm.

4.5. Choice of the analysis window

From the results of the preceding section, we can determine which windows $h(n)$ are suitable for the TD-PSOLA algorithm. The analysis window should be such that the short-time Fourier transform $X(\omega)$ is a reasonable estimate of the signal's spectral envelope. Two factors significantly influence the characteristics of $X(\omega)$: the type and the length of the analysis window $h(n)$.

As mentioned above, the length of the analysis window T is proportional to the local pitch-period $P(s)$, i.e., $T = \mu P(s)$. For standard window functions, the spectral resolution (mainlobe width) is inversely proportional to the window-length; the mainlobe width, in normalized frequency is $8\pi/T$ for the Hamming and the Hanning widows, and $12\pi/T$ for the Blackman window.

For $\mu = 2$ (wide-band analysis), the window's cutoff frequency ($4\pi/P(s)$ for a Hanning window) is larger than the spacing between the pitch-harmonics ($2\pi/P(s)$): the analysis window cannot resolve the individual pitch-harmonics. The short-time analysis spectrum $X(\omega)$ is a

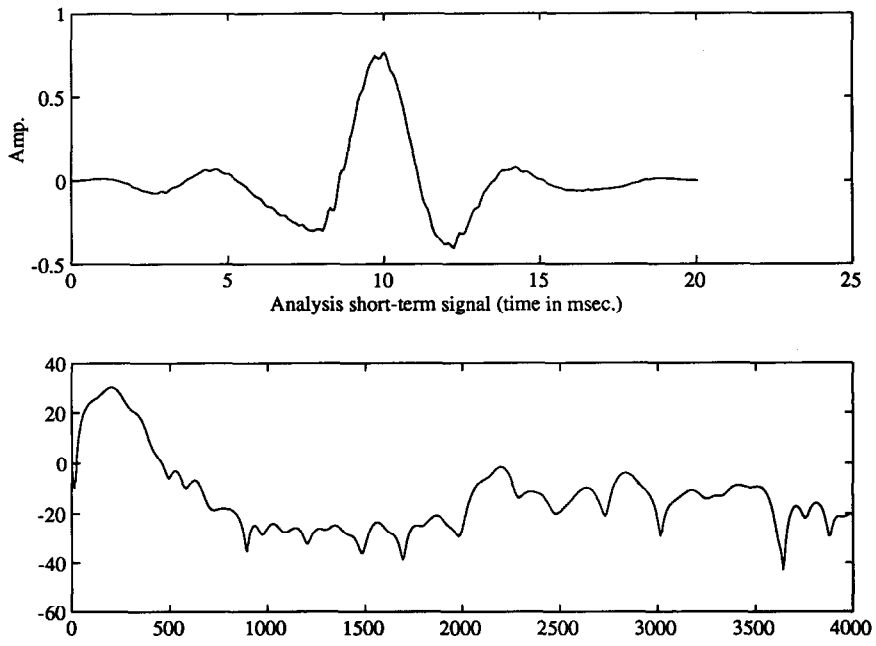

Fig. 11. Upper panel: short-time analysis signal. Lower panel: associated short-time analysis amplitude spectrum. The waveform is a vowel /i/, uttered by a male speaker (pitch frequency around 100 Hz).

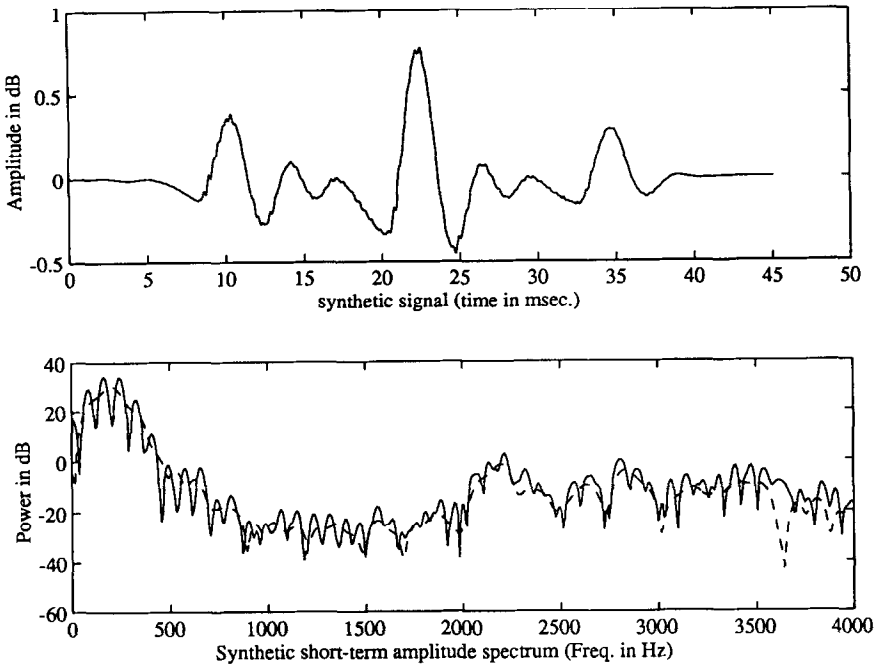

Fig. 12. The frequency resampling property. Upper panel: synthetic signal waveform. Lower panel: solid line: amplitude spectrum of the synthetic signal; dashed line: amplitude spectrum of the short-term analysis signal. The pitch-modification factor is equal to 0.8.

"smoothed" estimate of the speech signal spectrum envelope: the window mainlobe provides a means for interpolating between the pitch-harmonics.

By contrast, larger values of μ increase the window resolution and reveal the harmonic structure of $X(\omega)$, a property which is undesirable for the TD-PSOLA method: resampling $X(\omega)$ at the synthesis pitch-frequency $2\pi k/\beta P$ is likely to produce audible artifacts due to pitch-harmonic attenuation/cancellation.

Of course, the exact proportionality to the pitch-period is not mandatory; in real-time implementations, only a limited number of windows may be tabulated and stored in the computer memory, which avoids the need for computing new window coefficients for each pitch period.

The type of the analysis window used is also important: an excessive spectral leakage introduces undesirable timbre modifications by smearing the fine details of the formant structure. This remark excludes some peculiar windows introduced in previous contributions, such as the non-symmetric trapezoidal window proposed by Lukaszewickz and Karjalainen (1987) [9], or to a lesser extent, the non-symmetric raised cosine window proposed by Hamon (1988) in the original version of the TD-PSOLA algorithm.

The interpretation in terms of frequency resampling also explains some of the shortcomings of the TD-PSOLA method. It is clear that the spectral envelope implicitly used by the TD-PSOLA method to regenerate the synthesis complex pitch-harmonics does not exactly correspond to the "true" spectral envelope (i.e. $G(\omega)$ in the speech production model in 2.2) because of the frequency-domain convolution with $H(\omega)$.

This discrepancy becomes more acute for high-pitched voices, in which cases the analysis

[9] Note that a system very similar to the time-domain PSOLA is described in this contribution, but the authors failed to obtain high-quality speech synthesis because of the type of window function they used.

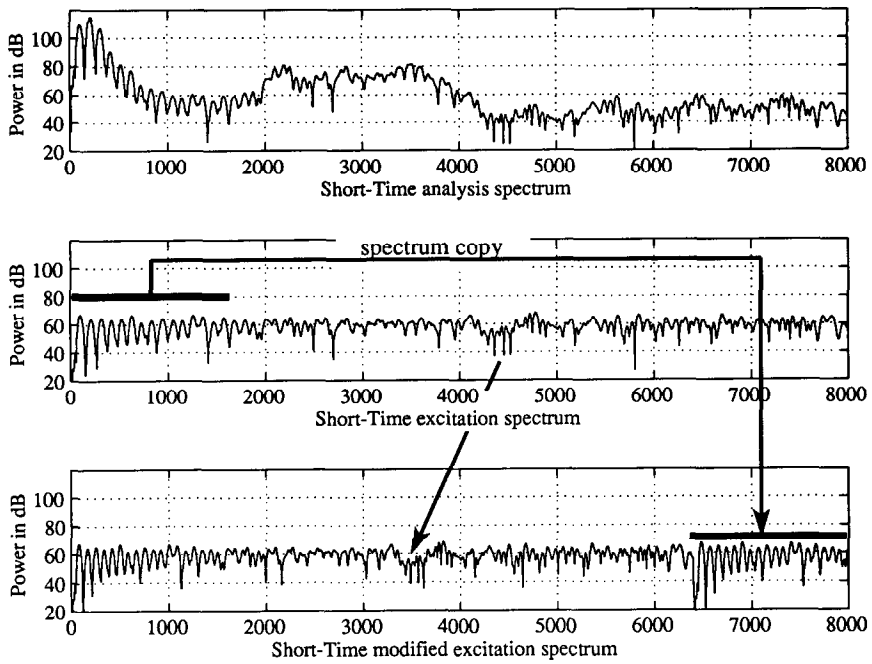

Fig. 13. Frequency-domain pitch-scale modification. Same short-time signal as in Fig. 10. Upper panel: original short-time amplitude spectrum. Middle panel: Excitation short-time amplitude spectrum (25 LPC coefficients). Lower panel: Modified short-time spectrum. Note that the upper frequencies have been regenerated using spectral copying. Also, the local deviations from a perfectly flattened spectrum are shifted in frequency.

windows are of shorter duration and therefore exhibit broader mainlobes. Typical errors are (i) formant bandwidth widening, especially for the first formant, and (ii) fusion of closely spaced formants, although this occurs less frequently. Fortunately, in general, these effects do not cause severe degradations of the quality of the speech output, at least when moderate pitch-scale modifications are undertaken.

4.6. Frequency-domain PSOLA and other variations

The Frequency-Domain PSOLA (FD-PSOLA) and the Linear-Predictive PSOLA (LP-PSOLA) approaches are theoretically more appropriate than the time-domain PSOLA method for pitch-scale modifications because they provide independent control over the spectral envelope of the synthesis signal. This point is particularly relevant in the context of speaker-characteristics modifications (Valbret et al., 1992).

FD-PSOLA. The FD-PSOLA technique is used only for pitch-scale modifications (Fig. 13) and differs from the TD-PSOLA algorithm in the definition of the short-time synthesis signals. Prior to overlap-add synthesis, each short-time analysis signal is modified; the modification is carried out in the frequency-domain on the short-time analysis spectrum. The algorithm is an exact replica of the frequency-domain resampling method described in Section 3.2.1: the spacing between the pitch-harmonics is scaled by a factor β_s, by linearly interpolating the real and imaginary parts of the complex spectra. As usual, when the pitch is lowered, the upper frequency band is regenerated either by spectral folding or copying. [10]

[10] Because these operations are performed at a pitch-synchronous rate, there is no need to adjust the phase spectrum with respect to the preceding short-time spectrum, as was the case when the analysis was not pitch-synchronous (Section 3.2.2).

LP-PSOLA. The Pitch-synchronous time-scale and pitch-scale modification scheme can also be easily embedded in residual-excited vocoders, yielding a method named Linear-Predictive PSOLA. Prior to any PSOLA processing, autoregressive models are fitted on the signal at a pitch-synchronous rate (an analysis filter is estimated for each analysis pitch-mark). Pitch-scale and time-scale modifications are then carried out on the residual signal, obtained by inverse filtering the original signal.

The output signal is finally obtained by filtering the modified residual with the time-varying synthesis filters, re-synchronized with the stream of synthesis pitch-marks.

The synthesis filters are either (i) simple copies of the analysis filters – in the standard elimination-duplication synthesis mode or (ii) interpolations between two successive analysis filters – in the more sophisticated interpolation synthesis mode. The methods used to improve the quality of RELP-synthesized speech (such as sample-by-sample interpolation of filter coefficients) can also be applied. Details are given in (Moulines and Charpentier, 1990; Valbret et al., 1992).

5. Other modification algorithms

This section presents several alternative techniques suggested mainly for time-scale modification of speech signals. Some of these techniques bear strong similarities to the two classes of methods presented above (STFT-based methods and Pitch-synchronous overlap-add methods), others propose slight improvements.

5.1. Methods based on the phase-vocoder

It was recognized as early as 1976 (Portnoff, 1981b) that time-scale modifications based on the phase-vocoder achieve good subjective results, but often introduce a variable degree of reverberation or *chorusing* [11] in the output signal, especially for large modification factors (above 2). Several methods have been proposed to counter this well-known artifact, some of which are briefly presented here.

5.1.1. Iterative reconstruction

As suggested in Section 2.1, the modified short-time Fourier transform does not necessarily correspond to any existing time-domain signal. The phase modification inherent to time-scaling (see Eq. (36)) does not preserve the phase coherence that exists in successive original short-time spectra. To avoid this problem, it was proposed (Hayes et al., 1980; Nawab et al., 1983; Griffin and Lim, 1984) that the phase information in the short-time spectra be discarded, and that the modified short-time Fourier transform be reconstructed from the knowledge of its magnitude only, using the large data redundancy in the STFT to make up for the loss of information. This idea, originally proposed in the field of image-processing, leads to iterative algorithms that converge to a local minimum of the distance in Eq. (5). Although the convergence of these algorithms has been proved in some cases (Griffin and Lim, 1984), it was remarked that the global minimum was not always reached.

These iterative reconstruction methods have been applied to the problem of time-scale modification (Nawab et al., 1983; Griffin and Lim, 1984; Roucos and Wilgus, 1985) and have been shown to improve significantly the quality of the modified speech signal. In particular, the reverberation/chorusing effect is significantly diminished. It was noted, however, that the convergence is usually quite slow (Roucos and Wilgus, 1985) and that the algorithms are extremely time-consuming.

5.1.2. Shape invariance

As observed in Section 3.1, time-scale modifications using the phase-vocoder are achieved by time-scaling the speech pitch-contour and the system amplitude function, and by letting the phases of the pitch-harmonics run free (Eq. (36)). As a result, the temporal aspect of the time-scaled signal differs significantly from that of the original signal because the local phase-relations between the pitch-harmonics are not preserved. This

[11] Chorusing refers to the subjective sensation that several persons are speaking at the same time.

phenomenon is quite visible in Fig. 9: during the first part of the signal, the temporal shape of the original and of the time-scaled signals are similar, the phase drift is still marginal. In the second part of the signal however (located right after an unvoiced segment), the temporal aspect of the time-scaled signal differs significantly from the temporal aspect of the original signal because the inter-harmonic phase relations have been lost.

A method has been suggested (Sylvestre and Kabal, 1992) in which the phase in each STFT channel is not allowed to run free, but is reset at each synthesis time-instant. As a result, the phase continuity between two successive short-time synthesis signals is no longer guaranteed. To recover phase continuity, a fixed phase offset is added to each channel, and the remaining phase discontinuity is exactly cancelled by slightly modifying the instantaneous frequency in each STFT channel. The modified signal is then obtained by concatenating the short-time signals, rather than overlap-adding them.

This algorithm guarantees that the phase relations between the pitch- harmonics in the vicinity of the synthesis time-instants are the same as in the original signal, in the vicinity of the analysis time-instants, up to a linear phase-shift. It is not sure, however, that the method brings significant improvement over the original phase-vocoder time-scaling method, in particular, with regard to the problem of unwanted reverberation/chorusing effect.

5.2. Time-domain methods

The time-scaled speech signals obtained by time-domain methods are generally of very high subjective quality and free of the reverberation/chorusing artifacts with which frequency-domain methods are often plagued. In particular, the inter-harmonic phase relations in the original signal are preserved in the time-scaled signal as seen in Fig. 9: the time-scaled signal keeps the same temporal aspect as the original signal. It should be noted, however, that these time-domain methods are based on the assumption that the speech signal is pseudo-periodic, which is no longer true in the case of several simultaneous speakers. The methods based on the phase-vocoder do not use this assumption: they merely require the signal to be composed of sinusoids with frequencies sufficiently far apart to be resolved by the analysis window. As a result, they do not break down even in the case of very complex signals (music (Dolson, 1986), etc).

5.2.1. The SOLA methods

In order to accelerate the iterative synthesis algorithm presented above (Section 5.1.1), Roucos and Wilgus (1985) have suggested using a better initial estimate of the phase, prior to any iteration. The initial estimate of the phase is chosen to be linear in frequency, thus corresponding to a simple time-delay. The value of the time-delay is obtained by maximizing the intercorrelation between the preceding short-time synthesis signal and the current one, the idea being to synchronize successive short-time signals. The authors found this initial estimate to be extremely good, to the point that further iterations of the reconstruction algorithm were not even necessary. When the iterative reconstruction procedure is discarded, it is possible to express the SOLA (Synchronous OverLap-Add) algorithm *in the time-domain*, without making use of Fourier transform.

The SOLA algorithm bears close similarities to the TD-PSOLA method, both using the same principle of overlap-adding short-time signals. They differ in the way the short-time signals are synchronized. The SOLA algorithm makes no explicit use of the pitch (although the cross-correlation maximization *actually is* pitch-detection), while the TD-PSOLA method runs in a pitch-synchronous manner.

Methods based on the same idea can be found in (Verhelst and Roelands, 1993; Hardam, 1990; Wayman and Wilson, 1988). Finally, it can be observed that, for moderate time-scale factors, the TD-PSOLA mapping function between the analysis time-instants $t_a(u)$ and the synthesis time-instants $t_s(u)$ (see Section 4.1.2) is generally a one-to-one mapping. For example, for a constant time-scale factor of $\beta = 1.1$, 9 analysis time-instants out of 10 are mapped to single synthesis time-instants, the 10th one being

mapped to two consecutive synthesis time-instants. In that case, time-domain algorithms (either TD-PSOLA of SOLA) obtain the time-scaled signal simply by copying the original signal from nine consecutive analysis frames, then duplicating the tenth frame and so on. There is obviously no need to calculate the pitch (nor the intercorrelation) for the nine frames where the short-time signals are simply copied. This observation inspired the algorithm described in (Laroche, 1993), in which segments of the signal are spliced or discarded according to the desired time-warping function, the splicing points being optimized by an intercorrelation maximization.

5.2.2. Parametric methods

This contribution has focused only on non-parametric methods: these methods rely heavily on the speech production model described in Section 2.2, but the parameters of this model are not estimated explicitly. Other techniques have been proposed in which the parameters of a speech production model are estimated, and explicitly used in the modification/synthesis stages.

The most straightforward of such approaches is the Linear Predictive Vocoder (Atal and Hanauer, 1971), in which voiced speech is modeled by convolving a periodic train of pulses by a time-varying IIR filter. This method has been used in many applications (e.g., text-to-speech systems), but is now abandoned because it fails to provide high-quality modifications.

Sinusoidal models represent a more promising approach. Sinusoidal models were initially proposed by Almeida and coworkers (Almeida and Silva, 1984, Marques and Almeida, 1989), and independently by McAulay and Quatieri (1986). Several variants have been described, see for example the paper by George and Smith (1992). In these models, voiced speech is represented as the sum of sinusoids with slowly time-varying amplitudes and instantaneous frequencies (which are not necessarily assumed to be harmonically related). The parameters of these models are obtained by use of short-time Fourier transform together with an appropriate peak-picking/tracking procedure.

The Harmonic + noise models are closely related to the sinusoidal model. They were originally proposed by Griffin and Lim (1988), mainly for low bit-rate speech coding applications, then further studied by Smith and Serra (Serra, 1989; Serra and Smith, 1990), Rodet and Depalle (Depalle, 1991; Poirot et al., 1988), Laroche, Moulines and Stylianou (Laroche et al., 1993a, 1993b) for speech modification applications. Here, the signal is decomposed as the sum of a deterministic component (sinusoidal with or without harmonically related frequencies) and a stochastic component, which accounts for (i) period-to-period fluctuations of the speech waveform in voiced segments and (ii) friction noise in unvoiced/mixed segments (note that the distinction between voiced/unvoiced segments is no longer binary).

The structures of the above models make it easy to devise rules for modifying the parameters in order to obtain desired time-scale/pitch-scale modifications. Because these models provide an accurate description of the speech signal, high quality modifications can be achieved (Quatieri and McAulay, 1986). A more thorough comparison with non-parametric approaches remains to be done.

6. Conclusions

In this contribution, several methods for time-scale and pitch-scale modification of speech signals have been presented in the common unifying framework of the short-time Fourier transform analysis/synthesis. A special emphasis has been put on the phase-vocoder and on the PSOLA methods, which are two typical examples of frequency-domain/time-domain techniques. Several variants of these schemes have also been described.

The main advantage of frequency-domain methods stems from the fact that the signal to be modified is not assumed to be quasi-periodic, which is useful when dealing with poly-pitch signals (e.g. simultaneous voices) and more generally for audio signals (e.g. music). However, as has been reported by most contributors, the modification often introduces an undesirable rever-

beration/chorusing effect. By contrast, time-domain methods rely heavily on the quasi-periodic assumption, but are free of reverberation/chorusing artifact, at least for moderate modification factors (between 0.5 and 2). Also, time-domain methods are generally more efficient in terms of computation load, especially in applications where the pitch can be pre-estimated (e.g. text-to-speech synthesis).

The many contributions that appeared in the last decade have made it possible to design high-quality prosodic transformation techniques. The degree of maturity of the techniques presented in this contribution should not hide the fact that many directions of research are still to be investigated.

1. As mentioned above, many variations/improvements of the basic methods presented in this paper have been proposed in the recent years. Although some of these methods are alleged to improve the quality of the modified signals, a thorough perceptual evaluation/comparison of frequency-domain, time-domain and parametric methods for time-scale and pitch-scale modifications remains to be done.
2. In the pitch-scale modification methods described above, only the harmonic structure is modified: the short-term speech spectrum envelope remains untouched. Especially when the pitch is raised even by moderate factors (e.g. 1.3–1.5), this results in an annoying alteration of the timbre: the modified voice becomes "thin" compared to the original voice. By analogy with the physics of voice production, it would seem reasonable to incorporate some degree of interaction between the fundamental frequency and the vocal tract transfer function. With this in mind, some of the voice transformation techniques described in this issue could be used to "learn" appropriate modifications of the vocal tract transfer function.

Acknowledgements

The authors would like to thank X. Rodet for his helpful comments.

References

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

J.B. Allen (1982), "Application of the short-time Fourier transform to speech processing and spectral analysis", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-82,* pp. 1012–1015.

L.B. Almeida and F.M. Silva (1984), "Variable-frequency synthesis: An improved harmonic coding scheme", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-84,* pp. 27.5.1–27.5.4.

B. Atal and S. Hanauer (1971), "Speech analysis and synthesis by linear prediction of the speech wave", *J. Acoust. Soc. Amer.,* Vol. 50, No. 2, pp. 637–655.

R.E. Crochiere (1980), "A weighted overlap-add method of short-time Fourier analysis/synthesis", *IEEE Trans. Acoust. Speech Signal Process.,* Vol. ASSP-28, No. 2, pp. 99–102.

R.E. Crochiere and L.R. Rabiner (1983), *Multirate Digital Signal Processing* (Prentice-Hall, Englewood Cliffs).

P. Depalle (1991), Analyse, modélisation et synthèse des sons basées sur le modèle source-filtre, PhD thesis, Université du Maine, Le Mans, France.

M. Dolson (1986), "The phase vocoder: A tutorial", *Computer Music J.,* Vol. 10, No. 4, pp. 14–27.

A. ElJaroudi (1991), "Discrete all pole modeling", *IEEE Trans. Acoust. Speech Signal Process.,* Vol. 39, No. 2, pp. 411–423.

A. ElJaroudi and J. Makhoul (1986), "All pole modeling for discrete spectra", *Proc. IEEE Workshop on Spectrum Estimation and Modeling, Boston,* pp. 29–32.

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

T. Galas and X. Rodet (1991), "Generalized functional approximation for source-filter system modeling", *Proc. Eurospeech, Genova,* pp. 1085–1088.

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 Engrg. Soc.,* Vol. 40, No. 6, pp. 497–516.

D.W. Griffin and J.S. Lim (1984), "Signal estimation from modified short-time fourier transform", *IEEE Trans. Acoust. Speech Signal Process.,* Vol. ASSP-32, No. 2, pp. 236–243.

D.W. Griffin and J.S. Lim (1988), "Multiband-excitation vocoder", *IEEE Trans. Acoust. Speech Signal Process.,* Vol. ASSP-36, No. 2, pp. 236–243.

C. Hamon (1988), Synthèse de parole par concaténation de formes d'ondes, Technical Report NT/LAA/TSS/RCP/359, CNET, Lannion, France.

E. Hardam (1990), "High quality time scale modification of speech signals using fast synchronized overlap add algorithms", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-90,* pp. 409–412.

M.H. Hayes, J.S. Lim and A.V. Oppenheim (1980), "Signal reconstruction from phase or magnitude", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, No. 6, pp. 672–680.

J. Laroche (1993), "Autocorrelation method for high quality time/pitch scaling", *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics.*

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

J. Laroche, Y. Stylianou and E. Moulines (1993b), "HNM: A simple, efficient harmonic plus noise model for speech", *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics.*

J.S. Lim and A.V. Oppenheim (1988), *Advanced Topics in Signal Processing* (Prentice Hall, Englewood Cliffs, NJ).

K. Lukaszewickz and M. Karjalainen (1987), "Microphonemic method of speech synthesis", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-87, Dallas,* pp. 1426–1429.

J.D. Markel and A.M. Gray (1976), *Linear Prediction of Speech* (Springer, Berlin).

J.S. Marques and L.B. Almeida (1989), "Frequency-varying sinusoidal modeling of speech", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 37, No. 5, pp. 763–765.

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

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.

S.H. Nawab and T.F. Quatieri (1988), "Short-time Fourier transform", in *Adanced Topics in Signal Processing,* ed. by J.S. Lim and A.V. Oppenheim (Prentice-Hall, Englewood Cliffs, NJ), Chapter 6.

S.H. Nawab, T. Quatieri and J.S. Lim (1983), "Signal reconstruction from short-time Fourier transform magnitude", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-31, No. 4, pp. 986–998.

A.V. Oppenheim and R.W. Schafer (1989), *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).

G. Poirot, X. Rodet and P. Depalle (1988), "Diphone sound synthesis based on spectral envelopes and harmonic/noise excitation functions", *Proc. Internat. Computer Music Conf., Köln.*

M.R. Portnoff (1980), "Time–frequency representation of digital signals and systems based on short-time Fourier analysis", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, pp. 55–69.

R. Portnoff (1981a), "Short-time Fourier analysis of sampled speech", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 29, No. 3, pp. 364–373.

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

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

S. Roucos and A.M. Wilgus (1985), "High quality time-scale modification of speech", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-85, Tampa, FL,* pp. 493–496.

S. Seneff (1982), "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-24, pp. 358–365.

X. Serra, A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition, PhD thesis, Stanford University, Stanford, CA, 1989, STAN-M-58.

X. Serra and J. Smith (1990), "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition", *Computer Music J.*, Vol. 14, No. 4, pp. 12–24.

B. Sylvestre and P. Kabal (1992), "Time-scale modification of speech using an incremental time-frequency approach with waveform structure compensation", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-92,* pp. 81–84.

H. Valbret, E. Moulines and J.P. Tubach (1992), "Voice transformation using PSOLA techniques", *Speech Communication,* Vol. 11, Nos. 2–3, pp. 175–187.

W. Verhelst and M. Roelands (1993), "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-93, Minneapolis,* pp. 554–557.

J.L. Wayman and D.L. Wilson (1988), "Some improvements on the sychronized-overlap-add method of time scale modification for use in real-time speech compression and noise filtering", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, No. 1, pp. 139–140.