Speech Communication 9 (1990) 453–467
North-Holland

PITCH-SYNCHRONOUS WAVEFORM PROCESSING TECHNIQUES FOR TEXT-TO-SPEECH SYNTHESIS USING DIPHONES

Eric MOULINES* and Francis CHARPENTIER**

Centre National d'Etudes des Télécommunications, Département Signal, 46 rue Barrault, F-75643 Paris Cédex, France

Received 1 August 1990

Abstract. We review in a common framework several algorithms that have been proposed recently, in order to improve the voice quality of a text-to-speech synthesis based on acoustical units concatenation (Charpentier and Moulines, 1988; Moulines and Charpentier, 1988; Hamon et al., 1989). These algorithms rely on a pitch-synchronous overlap-add (PSOLA) approach for modifying the speech prosody and concatenating speech waveforms. The modifications of the speech signal are performed either in the frequency domain (FD-PSOLA), using the Fast Fourier Transform, or directly in the time domain (TD-PSOLA), depending on the length of the window used in the synthesis process. The frequency domain approach is capable of a great flexibility in modifying the spectral characteristics of the speech signal, while the time domain approach provides very efficient solutions for the real time implementation of synthesis systems. We also discuss the different kinds of distortions involved in these different algorithms.

Zusammenfassung. Wir behandeln unter gemeinsamer Überschrift verschiedene Algorithmen, die in den letzten Jahren zur Verbesserung der Sprachqualität von Sprachvollsynthesesystemen vorgeschlagen wurden, die auf dem Prinzip der Verkettung akustischer Einheiten basieren. Kernstück dieser Algorithmen ist ein Verfahren der grundperiodensynchronen überlappenden Addition (PSOLA) zur Modifikation der Sprachprosodie und zur Verkettung von Sprachsignalsegmenten. Die Modifikationen des Sprachsignals erfolgen entweder im Frequenzbereich (FD-PSOLA), – hierbei wird die schnelle Fourier Transformation verwendet – oder direkt im Zeitbereich (TD-PSOLA), dies hängt von der Länge des jeweiligen Fensters ab, das bei der Synthese verwendet wird. Das Verfahren im Frequenzbereich erweist sich als besonders flexibel bei der Modifikation spektraler Eigenschaften des Sprachsignals, während das Verfahren im Zeitbereich sehr effiziente Lösungen für die Echtzeitimplementierung von Sprachsynthesesystemen bereitstellt. Wir diskutieren ebenfalls die verschiedenen Arten von Verzerrungen, die durch diese Algorithmen verursacht werden.

Résumé. Nous présentons différents algorithmes de génération du signal qui améliorent de façon significative la qualité sonore de systèmes de synthèse de parole à partir du texte par concaténation d'unités acoustiques (Moulines et Charpentier, 1988; Hamon et al., 1989). Ces algorithmes qui permettent de modifier les paramètres prosodiques et de concaténer les unités acoustiques sont fondés sur le principe de l'addition-recouvrement synchrone de la fréquence fondamentale de formes d'onde élémentaires (PSOLA). Les modifications du signal de parole sont réalisées soit dans le domaine fréquentiel (FD-PSOLA), soit directement dans le domaine temporel (TD-PSOLA), suivant la résolution spectrale et temporelle de la fenêtre utilisée lors du processus de synthèse. L'approche fréquentielle, plus complexe d'un point de vue algorithmique, permet de modifier avec une grande souplesse les caractéristiques spectrales du signal de parole. L'approche temporelle, d'une grande simplicité, est bien adaptée pour développer des systèmes de synthèse en temps réel. Nous discutons les différentes catégories de distorsions entraînées par ces algorithmes.

Keywords. Text-to-speech synthesis, voice quality, pitch-synchronous overlap-add (PSOLA).

1. Introduction

In this paper, we describe a family of methods for modifying the prosody of natural speech while retaining a high level of naturalness. These methods are used to improve the voice quality of text-to-speech systems based on the concatenation of elementary speech units, including diphones, demi-syllables, or non-uniform units as

* Present address: Ecole Nationale Supérieure des Télécommunications de Paris (Télécom Paris), Département Signal, 46 rue Barrault, 75643 Paris Cédex 13.

** Present address: Cap Sogeti Innovations, rue de Tocqueville, 75017 Paris.

proposed by Sagisaka (1988). Such concatenation-based synthesis systems require the use of rather large databases of acoustical units (corresponding for instance to 3 minutes of speech in the case of our French diphone system) and they generally rely on a coding algorithm to compress the size of the database. Therefore the synthesis stage generally involves two different processes:

- a decoding process: the waveform of the acoustical units must be reconstructed from their coded version;
- a concatenation process: the sequence of acoustical units must be concatenated after an appropriate modification of their intrinsic prosody.

The standard linear prediction method (LPC) integrates these two processes into a single step: since the fundamental frequency is an explicit parameter of the LPC synthesis model, the apparent pitch of the speaker can be altered by multiplying the derived estimate for fundamental frequency by a fixed factor prior to reconstruction of the synthetic speech. The temporal characteristics of the speech can also be altered by updating the parameters at the synthesizer at a different rate from the rate of extraction at the analyzer. However, such flexibility is counterbalanced by a limited voice quality: the oversimplification of the excitation signal – a repetitive pulse for voiced sound and random noise for unvoiced sound – causes the speech to sound indistinct or fuzzy. It also tends to smear abrupt changes present in the original speech. These problems can be partially solved through the use of more sophisticated excitation models, like "mixed" sources (periodic and noise-like excitations are mixed which have either time-invariant or time varying spectral shapes (Makhoul, 1978; Kang and Everett, 1985; Griffin and Lim, 1988)) or parametric glottal excitation (in these models, a parametric glottal waveform is estimated together with a model of the vocal tract transfer function (Hedelin, 1986; Lobo and Ainsworth, 1989)). However, these approaches fail to provide a completely natural speech quality.

In the methods we propose to improve voice quality, we explicitly separate the decoding and synthesis processes, thus synthesizing, as an intermediate step, the original waveforms of the acoustical units. The prosodic modifications are then performed directly on the speech signal, using a PSOLA waveform processing scheme. The fundamental frequency is an explicit parameter since the algorithm works at a pitch-synchronous rate and requires a preliminary pitch period labelling of the input waveforms. In particular, the synthesis process is synchronized with the synthesis pitch period, so that the algorithm simultaneously controls the value of the synthesized pitch and the duration of the synthesized signal.

In this paper, we first present the common PSOLA framework and two spectral interpretations of the synthesis process. We describe the time axis warping mechanisms to obtain time-scale and pitch-scale modifications. Then we successively analyze the time-domain and frequency domain algorithms, in relation with narrow-band and wide-band conditions of spectral analysis, and we explain the acoustical distortions involved in each approach.

1. The Pitch-Synchronous Overlap-Add synthesis framework

The PSOLA synthesis scheme involves the three following steps: an analysis of the original speech waveform in order to produce an intermediate non-parametric representation of the signal, modifications brought to this intermediate representation, and finally the synthesis of the modified signal from the modified intermediate representation. We present these three steps, and then we detail the specific features of the time-scaling and pitch-scaling algorithms.

1.1. Pitch-synchronous analysis

The intermediate representation of the digitized speech waveform $x(n)$ consists of a sequence of short-term signals $x_m(n)$, obtained by multiplying the signal by a sequence of pitch-synchronous analysis windows $h_m(n)$:

$$x_m(n) = h_m(t_m - n)x(n).$$

The windows are centred around the successive instants t_m, called pitch-marks, which are set at a

pitch-synchronous rate on the voiced portions of the signal and at a constant rate on the unvoiced portions. The windows $h_m(n)$ are of Hanning type and they are always longer than one single pitch period, so that neighbouring short-term signals (ST-signals) always involve a certain overlap. Their lengths are usually set to be proportional to the local pitch period, with a proportionality factor μ ranging from $\mu = 2$, for short analysis windows, to $\mu = 4$, for longer ones. These values correspond to 50% and 75% window overlapping factors, respectively. The proportionality rule for window length can be expressed as follows:

$$h_m(n) = h(n/\mu P),$$

where $h(t)$ is the window with the length normalized to unity and P denotes the local pitch period.

1.2. Pitch-synchronous modifications

The stream of analysis ST-signals $x_m(n)$ is converted into a modified stream of synthesis ST-signals $\tilde{x}_q(n)$ synchronized on a new set of synthesis pitch-marks $\tilde{t}_q$. Such a conversion involves three basic operations: a modification of the number of ST-signals, a modification of the delays between the ST-signals, and possibly, a modification of the waveform of each individual ST-signal. The number of synthesis pitch-marks $\tilde{t}_q$ depends on the pitch-scale and time-scale modifications factors called β and γ, respectively. The delays $\tilde{t}_q - \tilde{t}_{q-1}$ between two successive pitch-marks must be equal to the local synthesis pitch period. The algorithm works out a mapping $\tilde{t}_q \rightarrow \tilde{t}_m$ between the synthesis and analysis pitch-marks, specifying which analysis ST-signal $x_m(n)$ is to be selected to produce any given synthesis ST-signal $\tilde{x}_q(n)$. In the Time-Domain PSOLA (TD-PSOLA) approach, the synthesis ST-signals are obtained by simply copying a version of the corresponding analysis signal, so that the algorithm consists in selecting a certain number of analysis ST-signals $x_m(n)$ and translating them by the sequence of delays $\delta_q = \tilde{t}_q - t_m$:

$$\tilde{x}_q(n) = x_m(n - \delta_q) = x_m(n + t_m - \tilde{t}_q).$$

In the Frequency-Domain PSOLA (FD-PSOLA) approach, the synthesis ST-signals are obtained by a frequency-domain transformation of the translated signal $x_m(n - \delta_q)$.

1.3. Pitch-synchronous overlap-add synthesis

Several overlap-add (OLA) synthesis procedures are available to obtain the final synthetic speech. For instance, the synthetic signal $\tilde{x}(n)$ can be obtained by means of the least-square overlap-add synthesis scheme (Griffin and Lim, 1984):

$$\tilde{x}(n) = \frac{\sum_q \alpha_q \tilde{x}_q(n) \tilde{h}_q(\tilde{t}_q - n)}{\sum_q \tilde{h}_q^2(\tilde{t}_q - n)},$$

where $\tilde{h}_q(n)$ denotes the sequence of synthesis windows. The additional normalization factor α_q is introduced to compensate for the energy modifications related to the pitch modification procedure. The spectral interpretation of this synthesis scheme is that it minimizes the quadratic error between the spectra of the synthesis ST-signals $\tilde{x}_q(n)$ and the corresponding short-time spectra of the synthetic speech $\tilde{x}(n)$. An alternative synthesis scheme is the simple overlap-add procedure (Allen, 1977):

$$\tilde{x}(n) = \frac{\sum_q \alpha_q \tilde{x}_q(n)}{\sum_q \tilde{h}_q(\tilde{t}_q - n)}.$$

As in the least-square synthesis formula, the denominator of this formula plays the role of a time variable normalization factor: it compensates for the energy modifications due to the variable overlap between the successive windows. Under narrow band conditions, this factor is nearly constant. Under wide band conditions, it can also be kept constant, in particular when the synthesis window length is chosen to be equal to twice the synthesis pitch period. In such cases, and assuming $\alpha_q = 1$, the synthesis formula is reduced to the simplified overlap-add scheme:

$$\tilde{x}(n) = \sum_q \tilde{x}_q(n).$$

In this last formula, the synthetic signal appears as a simple linear combination of windowed and translated versions of the original signal. All

the operations involved are linear except the windowing operation. Subsequently, when combining the PSOLA speech modification scheme with a linear filter such as an LPC-filter or a low-pass filter, the order of the operations cannot be changed without modifying the behaviour of the overall system.

1.4. Time-scale modifications

Time-scale modifications of speech can be performed either in combination with pitch-scaling or as a separate transformation in itself. In the latter case, no frequency-domain modifications are required and only the TD-PSOLA algorithm is used, independently of the size of the analysis window.

In the simple case where the time-scale modification factor γ is constant, the $\tilde{t}_q \rightarrow t_m$ pitch-mark mapping associates $\tilde{t}_q$ with the analysis pitch mark, t_m being the nearest to the instant $\gamma \tilde{t}_q$. When slowing down the speech signal, the pitch-mark mapping results in the repetition of several analysis ST-signals (Figure 1). Inversely, a selective elimination of the analysis ST-signals leads to an acceleration of the speech signal.

In the cases where the speech is sped up or slowed down by a factor 2, and when the analysis window length is equal to two pitch periods, the TD-PSOLA algorithm is analogous to the former Time-Domain Harmonic Scaling (TDHS) algorithm (Malah, 1979), in which triangular windows are used. There is a greater difference between these algorithms for other time-scaling factors since the TDHS scheme rather uses a pitch-proportional synthesis rate, while the PSOLA schemes uses a truly pitch-synchronous rate. The TD-PSOLA algorithm must also be compared to the SOLA algorithm proposed in Roucos and Wilgus (1985), which works in an asynchronous way at the analysis stage and uses an autocorrelation technique to resynchronize the synthesis ST-signals with the pitch period.

The acoustical distortions involved by the TD-PSOLA scheme are negligible for accelerating the speech rate and for slowing it down by moderate factors. However, when slowing down unvoiced portions by factors in the range of 2 and above, the regular repetition of unvoiced ST-signals introduces a short-term correlation in the synthesized signal, which is perceived as a tonal noise. A practical solution for a factor 2 is to reverse the time-axis of every repeated version of an ST-signal. Higher factors are generally not required in applications such as diphone synthesis. However, if necessary, it is possible (although costlier) to use an FD-PSOLA scheme in order to randomize the phase spectrum of repeated unvoiced ST-signals.

Much slighter tonal noise distortions may also be perceptible for voiced sounds such as voiced fricatives, since their spectrum usually combines

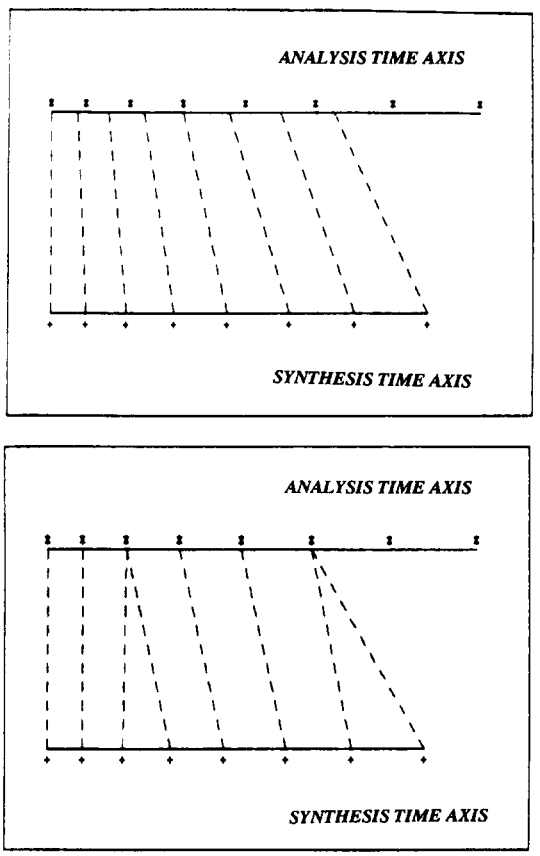

Fig. 1. Time-scale modifications using the TD-PSOLA algorithm. In this example, the speech is slowed down. The stars and plus signs represent the analysis and synthesis pitch-marks, respectively. In the first figure (above), the dashed lines represent the time-scale warping function between the analysis and synthesis time axes, corresponding to the desired time-scale modification. In the second figure (below), they represent the resulting pitch-mark mapping $\tilde{t}_q \rightarrow t_m$, in this case duplicating two analysis ST-signals out of six.

voiced and unvoiced frequency regions. Because of the presence of the voiced component, simple solutions such as time reversal are not feasible. A proper processing of such sounds would require an FD-PSOLA approach and a technique for identifying the distribution of voicing across the frequency axis.

1.5. Pitch-scale modifications

Pitch-scale modifications are slightly more complicated because they interfere with time-scale modifications. The simplest case is when the signal is to be simultaneously time and pitch-scaled by the same factors $\beta = \gamma$. There is then a one-to-one mapping between the synthesis and analysis pitch-marks: $\tilde{t}_m \rightarrow t_m$. This is illustrated in Figure 2 by the time warping function between the original time-scale and the second one below. But generally, independent time and pitch-scaling factors must be applied, so the mapping is not one-to-one and results in either duplication or elimination of some analysis ST-signals. As shown in Figure 2, this case can be seen as the combination of two transformations, the first one modifying the pitch and the time scale by the same factor β, the second one performing a compensatory time-scale modification by the factor γ/β. In fact, the two consecutive mappings that correspond to these two successive transformations combine into a single overall mapping, so that the time-scale and pitch-scale are performed simultaneously in one single step.

Finally, since pitch modification also involves the time-scaling mechanism, it should be noticed that slight tonal noises on voiced fricatives may occur when simultaneously raising the pitch and slowing down the signal since the inverse of the time-scaling factor γ/β may then become important.

2. TD-PSOLA approach

2.1. Interpretation of TD-PSOLA pitch modification

Some insight can be gained from studying the TD-PSOLA pitch modification of a somewhat idealized but realistic model of a stationary voiced sound consisting in the superposition of a deterministic periodic signal $d(m)$ and a zero-mean wide sense stationary (WSS) process $n(m)$ (Papoulis, 1984). The deterministic signal models the component which is repeated exactly from one pitch cycle to the next. The stochastic components take into account the variations from cycle to cycle (which can be observed on real speech): these variations can be attributed partly to irregularities in the vocal cord movement and partly to the turbulent airflow from the lungs during the open-glottis period in each pitch cycle (Kang and Everett, 1985; Griffin and Lim, 1988). This stochastic component can be predominant in some sounds (such as voiced fricative) or in certain conditions of phonation (i.e. mellow voice). In order to avoid artefacts, a pitch-modification algorithm should modify the period of the deterministic component without affecting too much

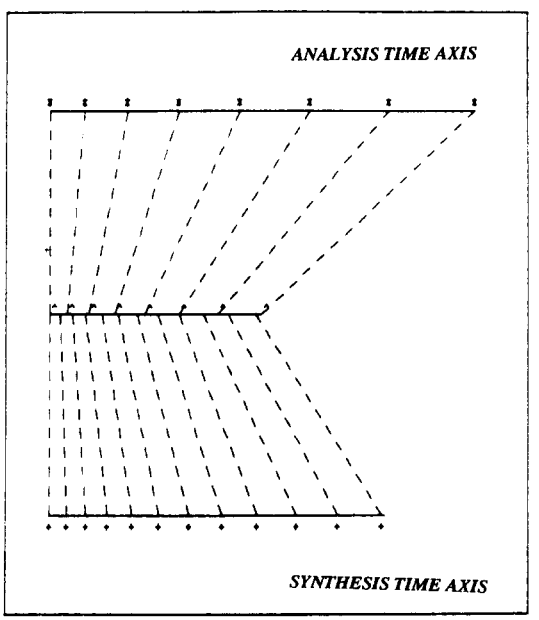

Fig. 2. Pitch-scale modifications using PSOLA algorithms. The stars, hat signs and plus signs indicate the analysis-marks, the modified pitch-marks on the virtual intermediate time axis, and the final synthesis ones. The dashed lines represent the time-scale warping functions between the successive time axes. A pitch and time-scale modification by the same factor is performed between the original and the intermediate time axes. A compensatory time-scale modification is performed between the intermediate and the synthesis time axes to achieve an overall independent time-scale modification.

the spectrum of the stochastic component. In the following, we assume that:

- the deterministic component is periodic with period P. The stream of analysis pitch-marks, which are set at a pitch-synchronous rate, are given by: $t_n = nP$,
- the pitch modification factor β is constant and equal to the time-scale modification factor so that there is a one-to-one mapping between analysis and synthesis pitch-marks: $\tilde{t} = n\beta P$,
- the simplified overlap and add scheme is employed to synthesize the signal. Moreover, the windows $h_n(m)$ are all equal to the same prototype window $h(m)$, which is chosen so as to comply with the following normalization condition: $\sum h(\tilde{t}_q - m) = 1 \ \forall m$.

With these assumptions, it can be demonstrated that the deterministic component of the synthetic signal is obtained by the periodization of a single prototype short-time signal $x_0(m)$ (which is equal to the product of the analysis window and the deterministic component of the analysis signal at the time origin) at the synthesis pitch period $\tilde{P} = \beta P$;

$$\tilde{d}(m) = \sum_s h(s\beta P - m)d(m - s\beta P)$$

$$= \sum_s x_0(m - s\beta P), \quad x_0(m) = h(-m)d(m).$$

The discrete Poisson formula, which relates the (infinite) sum of a signal on a periodic network to the (finite) sum of its Discrete Fourier Transform (DFT) coefficients at the associated harmonic frequencies, allows us to demonstrate that the Fourier series decomposition of the synthetic signal is given by:

$$d(m) = \frac{1}{\beta P}\sum_{k=0}^{\beta P-1} X_0(\tilde{\Omega}_k)\exp(j\tilde{\Omega}_k m), \quad \tilde{\Omega}_k = \frac{2\pi k}{\beta P}, \quad (1)$$

where $X_0(\omega)$ is the DFT of the prototype ST-signals. The complex amplitude of the synthetic signal pitch harmonics are equal to the DFT of the prototype ST-signal sampled at the corresponding frequencies $\tilde{\Omega}$. Therefore, the spectral envelope of the synthetic signal is identical to the short-term spectrum $X_0(\omega)$ of the original signal.

To determine the impact of the TD-PSOLA pitch modification algorithm on the stochastic component of our stationary voiced sound model, we first compute the autocorrelation sequence of the stochastic component of the synthetic signal:[1]

$$\tilde{n}(m) = \sum_s h(s\beta P - m)n(m - s\beta P + sP),$$

$$= \tilde{R}(\tau) = \frac{1}{\beta P}\sum_{k=0}^{\beta P-1} E(\tilde{n}(k)\tilde{n}(k-\tau)),$$

$$\tilde{R}(\tau) = \frac{1}{\beta P}\sum_m \varrho(\tau - m\beta P)R(\tau + m(1-\beta P)), \quad (2)$$

where $\varrho(\tau) = h(t)*h(-t)$ is the deterministic autocorrelation of the analysis window, and $\tilde{R}(\tau)$ is the autocorrelation of $n(m)$ (which is assumed to be known). Suppose, for example, that the process $n(m)$ is a white noise process with variance σ^2. The autocorrelation sequence of this process is equal to $R(\tau) = \sigma^2\delta(\tau)$ is the Kroenecker symbol. By applying formula (2), it can be shown that the autocorrelation sequence of the synthetic signal $\tilde{n}(m)$ exhibits a finite number of non-zero lags, which are disposed periodically, with a period equal to $(1 - \beta)P$:

$$\tilde{R}(k(1 - \beta P) = \varrho(kP)\sigma^2/\beta P.$$

Therefore, the synthetic signal $\tilde{n}(m)$ is no longer a white noise process: spurious correlation lags are introduced by the TD-PSOLA pitch modification algorithm in the autocorrelation sequence of the original process. More precisely, it can be seen in this particular case (but the result holds true in a more general context) that the autocorrelation sequence of the synthetic signal $\tilde{n}(m)$ is equal to the periodic duplication of the autocorrelation sequence of the original signal $n(m)$ with period $(1 - \beta)P$ (each shifted image is weighted by the deterministic autocorrelation of

[1] $\tilde{n}(m)$ is not a WSS process, but a WS cyclostationary process. This means that the autocorrelation sequence $R(t_1, t_2) = E(\tilde{n}(t_1)\tilde{n}(t_2))$ is not invariant on the diagonal of $t_1 - t_2$ plane, but is nevertheless periodic: $R(t_1 + m\beta P, t_2 + m\beta P) = R(t_1, t_2)$. However, there is a close relationship between WS stationary and WS cyclostationary processes: it can be demonstrated that if the time origin is subject to a random shift $\theta(t)$ that varies sufficiently slowly, then the resulting process $\tilde{n}(t) = \tilde{n}(t - \theta(t))$ is WS stationary (Papoulis, 1984). The autocorrelation sequence of this process is equal to the "time-average autocorrelation" of process $\tilde{n}(m)$.

the analysis window which is, for a large class of windows, a monotonously decreasing function equal to zero outside a compact subset, as the window is supposed to be of finite duration). Such a phenomenon can be seen as a time domain aliasing of the autocorrelation sequence. From formula (2), we can compute the Power Spectrum Density (PSD) of $\tilde{n}(m)$, which is the Fourier transform of its autocorrelation sequence:

$$\tilde{S}(\omega) = \frac{1}{\beta P^2} \sum_{k=0}^{P-1} |H(\omega(1-\beta) + \Omega_k)|^2 S(\beta\omega - \Omega_k), \qquad \Omega_k = \frac{2\pi k}{P}, \tag{3}$$

where $H(\omega)$ is the Fourier transform of the analysis window and $S(\omega)$ the PSD of $n(m)$. It can be seen from formula (3) that the Fourier transform of the analysis window $H(\omega)$ plays the role of the transfer function in the filtering of stochastic process. It is interesting to note that the DSP $\tilde{S}(\omega)$ of the synthetic signal at the frequency ω does not only depend on the value of the PSD $S(\omega)$ of the original signal at the same frequency: this is a direct consequence of the non-linear behaviour of TD-PSOLA pitch modifications.

Equations (2) and (3) show that the spectral envelope of the synthetic signal depends critically on the spectral resolution of the analysis window. We will study two extreme cases: for the first one, called narrow band (NB) conditions, the bandwidth of the analysis window $H(\omega)$ will be supposed to be small compared with the frequency resolution of interest, namely, the spacing between the pitch harmonics. For a "classical" window (Hanning, Hamming), which has low-pass characteristics, this condition is fulfilled as soon as the window length L is greater than 4 times the local pitch period: $L > 4P$. It will be shown that, in such a case, the TD-PSOLA pitch modification introduces a selective alteration of the amplitude of the pitch harmonics, which is a consequence of the discrepancy between the periodicity inherent in the synthesis ST-signal $\tilde{x}_q(n)$ and the synthesized pitch period. However, it preserves the sharp transitions of the PSD of the stochastic component $n(m)$ but can give rise to tonal noise for certain pitch modification factors.

In the second case, which will be referred to as wide band (WB) condition, the bandwidth of the analysis window $H(\omega)$ will be assumed to be greater than the local pitch period. This condition is satisfied for a window length less than 2 times the local pitch period $L < 2P$. In that case, the short-term prototype spectrum appears as a smoothed estimate of the true spectrum, the major discrepancies between these two spectrums lying in the bandwidth of formant resonances, which are broadened due to the convolution with $H(\omega)$ in the spectral domain. This convolution effect also smears out sharp transitions in PSD of $n(m)$, and, therefore involves some modification at the overlap between frequency regions dominated by harmonics of the fundamental and those dominated by noise-like energy.

2.2. Pitch modification under narrow-band conditions (Moulines, 1990)

(a) Deterministic signal. Equation (1) shows that the complex amplitude of a synthesis pitch harmonic is equal to the sample at the corresponding frequency of a short-term prototype spectrum $X_0(\omega)$ (derived from a ST-signal $x_0(m) = h(-m)x(m)$). Under narrow band analysis conditions, the bandwidth of the analysis window $H(\omega)$ is chosen to be the less than the local fundamental frequency. The prototype ST-spectrum $X_0(\omega)$ thus consists of the non-overlapping weighted images of $H(\omega)$, shifted to the original pitch harmonics frequencies Ω_k.

As a consequence, it appears that the amplitude of a synthesis pitch harmonic will be all the more affected as it departs from the original pitch harmonics. This behaviour is illustrated in Figure 3, where the prototype ST-spectrum is displayed along with the spectrum of the synthetic signal; the pitch modification factor is equal to $\beta = 3/2$. Two out of three pitch harmonics are attenuated, whereas the third one passes through TD-PSOLA pitch modification without amplitude alteration since its frequency corresponds exactly to a pitch harmonic of the original signal. The worst case occurs when the synthesis harmonic lies right in the middle of two neighbouring original pitch harmonics: in which case, the complex harmonic amplitude is almost zero. If F_0 is the

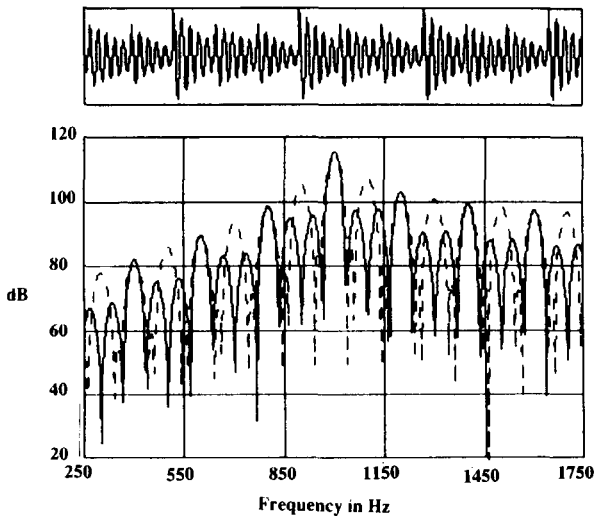

Fig. 3. TD-PSOLA pitch modification under narrow band conditions. The short-term spectrum of the prototype signal (dashed line) provides the envelope of the synthetic signal (solid line). The synthetic signal is displayed above the spectral representation. The original signal is the response of a second order recursive filter to a periodic unit sample train. The pitch modification factor is equal to $\beta = 3/2$. The length of the analysis window (Hanning) has been set to 5 times the local pitch period.

original fundamental frequency and if the pitch-modification factor satisfies the constraint $1/2 < \beta < 3/2$, the frequency zones of maximum attenuation lie around the multiples of the following frequency:

$$f_{\text{att}} = \frac{F_0}{2} \frac{\beta}{|\beta - 1|}.$$

Such attenuations can be observed even for very mild modification factors such as $\beta = 1.05$, since for a pitch value of $F_0 = 100$ Hz, the attenuation zone then lies in the frequency zone around $f_{\text{att}} = 1000$ Hz.

These kinds of distortions can be interpreted as a reverberation of the desired pitch-modified signal, involving echoes with a time variable delay equal to the original pitch period (Moulines, 1990). This is confirmed by informal listening tests which prove that TD-PSOLA pitch modification leads to reverberant-sounding distortion.

(b) Stochastic signal. For moderate pitch modification factors $0 < \beta < 2$, it can be shown that the PSD of the synthetic signal will present deep valleys of fixed width ω_w centered on a set of points ω_k periodically spaced along the frequency axis:

$$\hat{\omega}_k = \frac{(2k+1)\pi}{P|\beta - 1|}, \qquad \omega_w = \frac{2}{|\beta - 1|}\left(\frac{\pi}{P} - \omega_h\right),$$

where ω_h is the cutoff frequency of $H(\omega)$. Moreover, if we assume that the PSD $S(\omega)$ is a slowly varying function of frequency and that it is approximately constant over the bandwidth of the analysis window $H(\omega)$, it appears that the PSD of the synthetic signal consists of the shifted images of $H(\omega)$, weighted by the PSD of the original signal

$$\tilde{S}(\omega) = \begin{cases} \dfrac{1}{\beta P} S(\hat{\Omega}_k) |H((1-\beta)(\omega - \hat{\Omega}_k))|^2 \\ \times |\omega - \hat{\Omega}_k| < \dfrac{\omega_h}{|\beta - 1|}, \quad \hat{\Omega}_k = \dfrac{2\pi k}{|\beta - 1| P}, \\ 0 \quad \text{otherwise.} \end{cases}$$

This relation shows that TD-PSOLA formula introduces a "pseudo-periodic" structure: the synthetic signal can be seen as the output of a comb filter whose frequency response consists of non-overlapping images of $H(\omega)$, shifted to the frequency $2\pi k/|\beta - 1|P$. This behaviour is illustrated in Figure 4, which shows the Bartlett estimator of the PSD of a synthetic signal produced by TD-PSOLA pitch modification of a band-pass noise: the pseudo-harmonic structure is clearly exemplified. From informal listening tests, it appears that the pseudo-periodic structure manifests itself as tonal noise (whistling).

2.3. Pitch modification under wide-band conditions (Lukaszewickz and Karjalainen, 1987; Hamon et al., 1989)

(a) Deterministic signal. For wide band analysis, the bandwidth of the analysis window $H(\omega)$ is chosen to be several times greater than the fundamental frequency (for a "classical" window with length equal to two times the local pitch period, the bandwidth of $H(\omega)$ is 4 greater than the fundamental frequency). Because the bandwidth of $H(\omega)$ is greater than the frequency

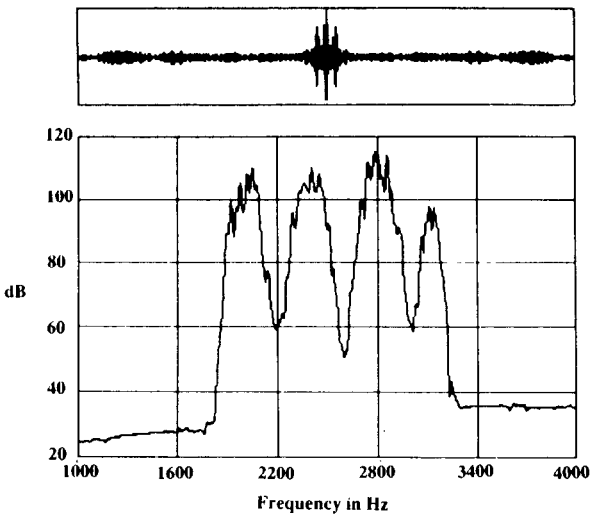

Fig. 4. TD-PSOLA under narrow band conditions – stochastic signal. The autocorrelation sequence of the synthetic sequence is displayed above the spectral representation. The original signal is a band pass noise obtained by filtering a white noise process by an 8th order elliptic band-pass filter (with cutoff frequencies equal to 2000 Hz and 3000 Hz respectively). The autocorrelation sequence is estimated by using the standard biased estimator over 5000 samples. The Bartlett estimator of the PSD is computed over 512 correlation lags. The underlying pitch period is equal to 100 Hz. The pitch modification β is equal to 3/4. The window length is set to 3 times the local pitch period (between WB and NB conditions). As predicted, zeros appear periodically every 400 Hz.

spacing between pitch harmonics, the window does not have sufficient frequency resolution to resolve the individual pitch harmonics and thus provides a means for interpolating between the pitch harmonics to obtain a "smoothed" estimate of the "target" spectral envelope (Portnoff, 1981). However, this "smoothed" estimate (and subsequently the pitch harmonics amplitude of the synthetic signal), involves a certain error with respect to the desired spectral envelope, due to the convolution with $H(\omega)$ in the spectral domain.

The major distortion appears to be a broadening of the formant resonances: because the bandwidth of a formant is usually much less than the bandwidth of the analysis window, $|H(\omega)|$, the shape of $X_0(\omega)$ "formant" peaks depends primarily on the analysis window: the bandwidths of these peaks do not provide good estimates of the "true" formant bandwidth. The problem appears to be more severe when the original pitch increases, as the spectral resolution of $H(\omega)$ has to be reduced to meet WB analysis conditions. In some adverse cases, a fusion of closely spaced formant peaks can even be observed. Hopefully, from a perceptual point of view, this mismatch of bandwidth is not too severe most of the time, since the differential limen for the perception of formant bandwidths is very high (psychoacoustical measures show that this limen is about 40% for steady vowel and should be much larger for continuous speech (Flanagan, 1972; Ghitza, 1986)).

The effect of TD-PSOLA on the short-time phase spectrum is difficult to analyse. Even in the idealized stationary periodic case, the phase of a synthetic signal harmonic cannot be simply deduced from the original signal phase distribution. To go further, we adopt a simplified model of voiced speech production; we assume that a stationary voiced speech is the output of a linear time in invariant all-pole filter $T(z) = 1/A(z)$ driven by a periodic train of unit samples: $v(n) = \sum\delta(m - D + rP)$. Since the analysis pitch marks t_n are set on samples which are integer multiples of the period $P(t_n = nP)$, D corresponds to the number of samples of the pitch mark t_n from the unit sample arriving most recently before. If we suppose that the analysis window is even-symmetric $h(m) = h(-m)$ (which implies that its Fourier transform is real), we can demonstrate that the synthetic signal phase distribution critically depends upon the delay D (Moulines, 1990). More precisely, if we assume that:

- the bandwidth of each resonance (or *formant*) is much less than the bandwidth of the analysis window $H(\omega)$;
- the frequency spacing between formants is greater than the bandwidth of $H(\omega)$, the phase of a synthetic signal pitch-harmonic can be shown to be equal to:

$$\theta(\tilde{\Omega}_l) \approx \theta_0(\tilde{\Omega}_l) + (\omega_i - \tilde{\Omega}_l)D, \quad 0 \leq D < P,$$

where $\tilde{\Omega}_l$ is the synthetic harmonic frequency, ω_i is the normalized frequency of the formant peak that is the nearest to the synthetic harmonic, and $\theta_0(\tilde{\Omega}_l)$ is the phase transfer function of $T(z)$. This result suggests that minimum phase distortion is achieved when the window is synchronized with the main excitation of the vocal tract within the

pitch period, namely the instant of glottal closure. Experimentally, when the window centre is shifted with respect to the presumed glottal closure instant, the synthetic speech is first altered and then begins to sound hoarse when the shift exceeds 30% of the pitch period. This degradation however cannot be simply attributed to the synthetic signal phase distribution. It probably also results from distortions of formant amplitudes since the short-term spectrum of the prototype signal may vary a lot with the window position. If we adopt the simplified model of voiced speech production specified above, it appears that the magnitude of $X_0(\omega)$, for a given frequency ω, is a function of the delay D between the pitch marks and the unit-samples of the excitation signal. More precisely, the short-term prototype spectrum $X_0(D, \omega)$ (where D explicitly denotes this time dependence) behaves in time as a decaying exponential with a time constant determined by the bandwidth of the nearest formant in the neighbourhood of ω. In practice, this means that the estimated formant amplitudes (used implicitly to reconstruct the synthesis signal) depend on the window position with respect to the instant of the main excitation of the vocal tract. Experimentally, we have observed that an improper synchronization of the synthesis window more strongly affects the amplitudes of high-frequency formants, which have larger bandwidths (and thus have a rapid decay).

2.4. The LP-PSOLA approach (Moulines and Charpentier, 1988; Moulines, 1990)

In fact, it is possible to combine the LD-PSOLA approach and LPC techniques using either the prediction error signal itself or a low bit rate encoded version of it (MPLPC, CELP). Such low-bit rate coding algorithms are useful in the context of concatenative synthesis systems since they provide a means for reducing the amount of memory required to store the speech databases, as mentioned in the introduction. A straightforward method is to perform the TD-PSOLA modifications of speech after the LPC decoding process, thus synthesizing the original waveforms as an intermediate step. However, it is also possible to perform the TD-PSOLA modifications directly on the excitation signal of the LP filter, thus exchanging the LPC filter and the TD-PSOLA synthesis scheme. Such a modification of the synthesis structure defines the *Linear Predictive* PSOLA (LP-PSOLA) approach. The TD-PSOLA and LP-PSOLA are not equivalent since, as pointed out above, it is not possible to exchange the order of the LPC-filtering operation and of the windowing operations inherent in the TD-PSOLA process. It should be noted that the LP-PSOLA method can also be seen as an extension of previous cut-and-splice methods, where the cutting of individual pitch periods of the residual is replaced by a smooth windowing operation, allowing inter-period overlap (Caspers and Atal, 1983; Van Hemert, 1984; Varga and Fallside, 1987).

A theoretical advantage of the LP-PSOLA approach is the better resolution that can be achieved with parametric estimation techniques compared to the one achieved with the short-term Fourier transform implicitly used in the TD-PSOLA approach: the use of a slowly time-varying LP filter to model the short-time spectral envelope of the quasi-stationary speech signal has proven to be a very successful approach in many areas of speech processing. Moreover, the splitting of the speech spectrum into a source and a filter component provides an additional degree of freedom that can be exploited in two ways. First, a different time frame for spectral envelope estimation than for the prosodic modifications can be used. This flexibility allow us, for example, to adopt narrow bandwidth conditions for spectral envelope estimation (to get a smoothed estimation of the vocal tract transfer function and the spectral characteristics of the glottal pulse) or, on the contrary, to perform the estimation on short time frame within the interval of glottal closure (to obtain an accurate approximation of the vocal tract transmission characteristics). Second, joint modifications of the pitch and of the spectral envelope can be performed: this ability, which is not fully exploited yet, is important since there is some practical evidence showing that the spectral envelope changes as the pitch varies due partly to articulatory adjustments and partly to modifications of the glottal waveform.

The improved spectral modelling can be

exploited since the TD-PSOLA scheme involves less distortion when working on the residual waveform than on the speech signal itself. Indeed, the spectrum of the prediction error signal has an approximately flat spectral envelope, with slight variations as the LP-filter is generally not capable of removing the speech resonant and anti-resonant frequency components completely. These variations may be approximated by an amplitude spectrum with broadened spectral peaks. The bandwidths of such broadened resonances are much wider than the main lobe of the synthesis window used in the wide-band TD-PSOLA scheme. Consequently, the LP-PSOLA algorithm is capable of maintaining the flatness of the modified residual spectrum and of reproducing the original spectral envelope in the synthesized speech. Furthermore, it can preserve the wide-band spectral deviations inherent to the residual signal (in that respect, the method is similar to the FD-PSOLA approach using the *elimination-repetition* technique to preserve the spectral deviations).

Finally, the different behaviour of the TD-PSOLA and the LP-PSOLA pitch modification algorithm is illustrated in Figure 5, for a vowel /i/ pronounced by a female speaker. As we can see, TD-PSOLA pitch modification introduces a broadening of formant bandwidths. This effect, whose impact can be observed both on the synthetic signal and on its short-term spectrum, is particularly evident for the first two formants (the corresponding "original" formants have narrow bandwidth in accordance with the acoustic theory of speech production). As we can see, the behaviour of the LP-PSOLA pitch modification method is different. The synthetic formant bandwidths, derived from the LP-spectrum envelope which results from a WCA analysis (Weighted Covariance Analysis (Miyoshi et al., 1987)) are more compatible with those predicted by theory. Surprisingly, although the synthesized signals obtained by both methods look quite different, it is difficult to distinguish them from an auditory point of view.

3. FD-PSOLA approach (Charpentier and Moulines, 1988)

When using the TD-PSOLA pitch modification scheme under NB analysis conditions, we saw that the mismatch between the residual periodicity of the ST-signals and the synthesized pitch value in-

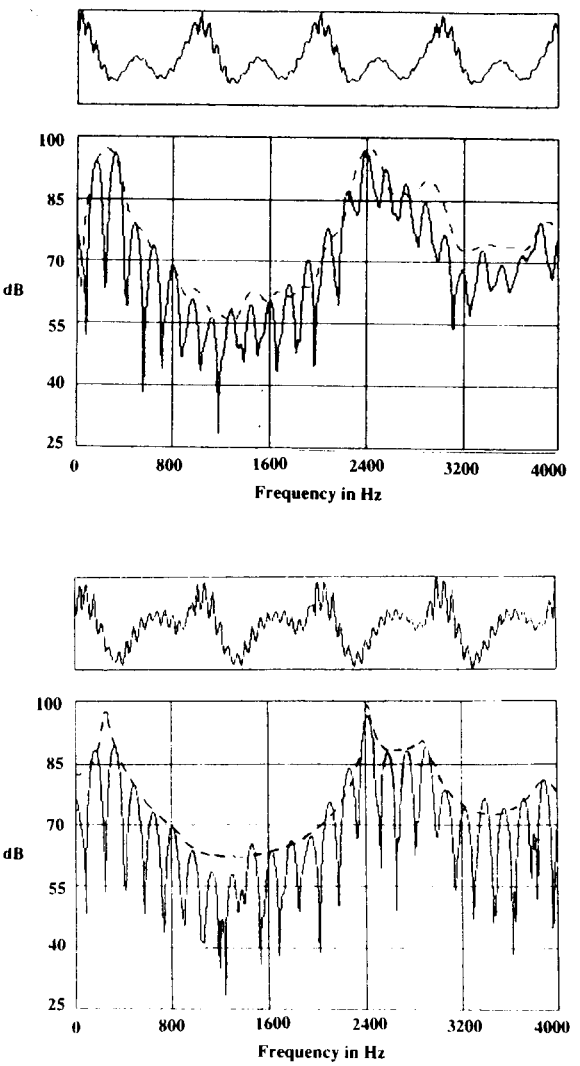

Fig. 5. Pitch modifications under wide band conditions. Pitch modification of a natural vowel /i/ uttered by a female speaker (β = 3/2, f_0 = 260 Hz). The synthetic signals obtained by the two methods are displayed along with their short-term spectrum using a 25 ms Hanning window.
(a) TD-PSOLA method: the dashed line indicates the magnitude spectrum of the ST-synthesis signal which corresponds to the implicit spectral envelope; (b) LP-PSOLA method: the dashed line indicates the LP-spectrum envelope.

troduces selective attenuation of frequency bands, leading to reverberant-sounding distortions. To avoid this problem, an attractive solution is to perform appropriate spectral modifications on each voiced analysis ST-signal before feeding it to the PSOLA synthesis scheme. This solution will be referred to as *Frequency Domain PSOLA*, since this synthesis scheme involves modifications of the signal in the frequency domain. In this respect, the FD-PSOLA method can be compared with previously proposed methods using the discrete Fourier transform spectrum of the input signal as intermediate data on which modifications are performed. In this category fall the digital phase vocoder (which has been widely used in speech processing and in computer music applications) (Seneff 1982; Moorer, 1978) and the asynchronous Overlap Add method (Allen, 1974).

As for the PSOLA general gramework, these modifications can also be divided into an analysis, a modification and a synthesis step.

(a) Frequency-domain analysis. The analysis process can be divided into three steps. First, a complex short-term spectrum is computed which corresponds to the Discrete Fourier Transform (DFT) of the ST-signal with the time origin set to coincide with the analysis pitch-mark t_n. Second, a short-term spectral envelope is estimated via standard parametric modelling techniques (such as linear prediction), or via non-parametric homomorphic (cepstral) analysis. Finally, a flattened version of the ST-spectrum is derived (the "source component") by dividing the ST complex spectrum by the global spectral envelope. The spectral representation therefore corresponds to the classical *source-filter* model, consisting of a spectral envelope and a spectral representation of the source.

(b) Frequency-domain modifications. To obtain a pitch-modified version of the spectrum, the source component of the spectral representation is modified so that the spacing between the pitch harmonics is equal to the new fundamental frequency. We present here two different methods to do this (Charpentier, 1988).

The first method is the spectral *compression-expansion* technique illustrated in Figure 5. The frequency axis of the spectrum is linearly warped by the pitch-modifications factor β. To do this, the real and imaginary parts of the original spectrum are linearly resampled to obtain the modified DFT coefficients: the $\tilde{k}$-th coefficient of the synthetic ST complex spectrum $\tilde{S}_{\tilde{k}}$ is derived from the "virtual" k/β-th coefficient of the original complex spectrum. Since k/β is generally not integer, we perform a simple linear interpolation:

$$\tilde{S}_k = (1-\alpha)S_{\bar{k}} + \alpha S_{\bar{k}+1}, \quad \bar{k} = \left[\frac{k}{\beta}\right], \quad \alpha = \tilde{k} - \bar{k},$$

where $[x]$ is the truncation operator which converts a real to its corresponding integer value. As schematized in Figure 6, this method introduces a certain kind of spectral distortion: since it implicitly maintains a one-to-one mapping between the original and the synthesis pitch harmonics, it modifies the fine details of the spectrum by carrying its local properties, such as harmonics phases and amplitudes, or the local voiced/unvoiced feature, from one zone of the spectrum to another. A related problem also occurs when decreasing the pitch, since an empty spectral zone appears in the high frequencies, requiring techniques for generating an acceptable spectral distribution in those regions (two methods, which lead to approximately equal qualities, have been investigated: copying the lower part of the spectrum to the upper part or folding the high frequencies (Seneff, 1982; Charpentier and Moulines, 1988)). Hopefully, in spite of these distortions, the method is capable of maintaining very good speech quality.

In fact, the *compression-expansion* pitch modification method works by resampling the origional source spectrum at a modified set of frequency points, and it can be seen as a frequency-domain equivalent of a time-domain method developed in the context of RELP synthesis: the residual resampling method. The advantage of the FD-PSOLA approach is that it allows a time-variable resampling rate (and thus very flexible pitch modifications), difficult to realize in the time-domain.

An alternative method is the harmonics *elimination-repetition* technique, also illustrated in Figure 6. This method attempts to overcome the dis-

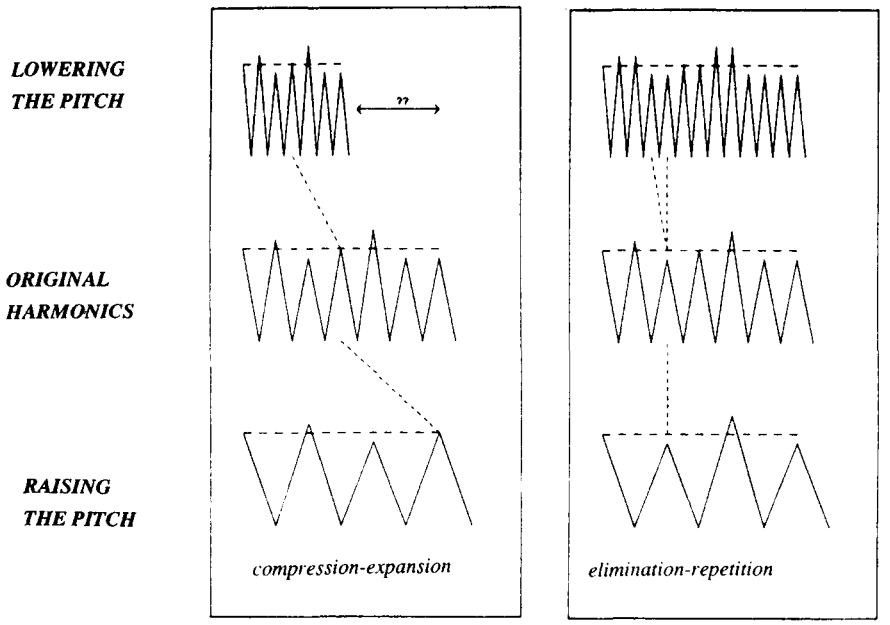

Fig. 6. Pitch modification using the FD-PSOLA algorithm and two different spectral resampling methods: the *compression-expansion* method (left) and the *elimination-repetition* method (right). The spectral deviations with respect to an ideally flat source spectrum are enhanced by the horizontal dashed lines. The other dashed lines illustrate the correspondence between the synthesis and original pitch harmonics.

tortions of the previous method by preserving the local spectral properties of the spectrum. The method has a similar organisation in the frequency domain as the PSOLA pitch-scaling methods in the time-domain. When lowering the pitch, the algorithm inserts new harmonics by repeating original ones while reducing the spacing between them. When raising the pitch, harmonics are eliminated and the inter-harmonics spacing is increased. This method has some drawbacks: it requires a precise estimation of the pitch value and it modifies the phase coherence between successive harmonics that is implicitly maintained by the *compression-expansion* algorithm. In order to ensure good quality, the algorithm must compensate for the linear component of the phase spectrum, or equivalently, the original pitch-marks must be set at locations corresponding to the maximal source excitation. In fact, the flexibility of the FD-PSOLA approach allows to combine the two methods into a mixed method, using the *compression-expansion* scheme in the low frequencies in order to preserve the phase coherence of the first harmonics, and the *elimination-repetition* scheme in the high frequencies in order to avoid the need of high frequency generation.

Finally, we should mention that the FD-PSOLA approach can support a wide variety of modifications such as:

- zero-phasing the low frequencies: to obtain a homogeneous phase distribution, while avoiding the buzziness due to a zero-phase spectrum in the high frequencies (Charpentier and Moulines, 1988);
- modifications of the voice quality: this can be obtained through linear or non-linear modifications of the envelope component or selective amplification or attenuation of individual pitch harmonics (Allen, 1977; Richards, 1974);
- sampling rate modifications: this approach provides an alternative to classical filtering techniques and is very flexible for modifying the sampling rate by arbitrary ratios. However, the pitch synchronization is not crucial for this application, and an asynchrnous OLA scheme can be utilized (Crochiere and Rabiner, 1981).

(c) Short-term signal synthesis. Finally, the modified representation is converted back to a synthetic complex spectrum, and the synthesis ST-signal is obtained by inverse Fourier Transform.

Both interpretations of the OLA synthesis schemes are valid in the case of the FD-PSOLA approach. The least-square synthesis scheme will perform a good match between the modified ST-spectra and the spectrum of the synthesized signal, since they will have a common fine spectral structure. On the other hand, the simplified OLA method will also perform a sound modification, since the amplitude of the modified ST-spectra at the synthesis harmonic frequencies have the right target values, i.e. the values of the spectral envelope.

4. Concluding remarks

The PSOLA algorithms in this paper provide the means for modifying the prosodic parameters of natural speech, such as pitch, duration and energy, while maintaining a very good voice quality in the transformed speech. The FD-PSOLA algorithm requires a high computational power (about 5Mflop/s.) and a high memory size for the digitized diphone dictionaries (about 5Mbytes). However, it provides a flexible laboratory tool since it allows a fine control of the speech spectrum. The TD-PSOLA algorithm is computationally very efficient and it can be combined in a flexible manner with speech compression techniques in order to reduce memory requirements. The LP-PSOLA algorithm can be seen as an optimized combination of the TD-PSOLA algorithm, and of specific LPC coding technique (such as MPLPC or CELP).

The speech quality gain brought by these algorithms has been evaluated through a formal test on French synthesized speech using diphones. This test was performed on 16 subjects and on 10 sentences comparing an LPC synthesizer with improved excitation and early versions of the FD-PSOLA, TD-PSOLA and LP-PSOLA algorithms. The four systems were compared two by two in A–B and B–A pairs for preference. The results have shown that all three algorithms perform much better than LPC synthesis and that they are relatively equivalent among themselves.

Current developments include real-time implementation of a multilingual diphone synthesis system in two different configurations: using the TD-PSOLA algorithm on a personal computer using an Intel 80386 processor and using the LP-PSOLA algorithm (also called PSOLA-MPLPC, because a low bit-rate MPLPC is used to code the diphone database) on a Texas TMS320C25 based digital signal processing board.

References

J.B. Allen (1977), "Short-term spectral analysis, synthesis, and modification by discrete Fourier Transform", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP-25, pp. 235–238.

J.B. Allen and L.R. Rabiner (1977), "A unified approach to short-time Fourier analysis and synthesis", *Proc. IEEE*, Vol. 65, No. 11, pp. 1558–1564.

B.E. Caspers and B.S. Atal (1983), "Changing pitch and duration in LPC synthesized speech using multipulse excitation", *J. Acoust. Soc. Am.*, pp. 73, S5(A).

F. Charpentier (1988), "Traitement de la parole par analyse-synthèse de Fourier: Application à la synthèse par diphones", *Doctoral thesis, Ecole Nationale Supérieure des Télécommunications.*

F. Charpentier and E. Moulines (1988), "Text-to-speech algorithms based on FFT synthesis", *Proc. Int. Conf. Acoust., Speech, Signal Proc., New York*, pp. 667–670.

R.E. Crochiere and L.R. Rabiner (1981), "Interpolation and decimation of digital signals - A tutorial review", *Proc. IEEE*, Vol. 69, No. 3, pp. 300–341.

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

O. Ghitza (1986), "Speech analysis and synthesis based on matching the synthesized and original representation in the auditory nerve level", *IEEE Int. Conf. Acoust., Speech, Signal Proc., Tokyo*, pp. 1995–1998.

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

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

C. Hamon, E. Moulines and F. Charpentier (1989), "A diphone synthesis system based on time-domain modifications of speech", *Proc. Int. Conf. Acoust., Speech, Signal Proc., Glasgow*, pp. 238–241.

P. Hedelin (1986), "Glottal LPC vocoding", *Int. Conf. Acoust., Speech, Signal Proc., Tokyo*, pp. 465–468.

J.P. van Hemert (1984), "Multipulse excitation: The possibilities and restriction of a new speech synthesizer",

IPO Progress Report No. 19, pp. 20–24.

G.S. Kang and S. Everett (1985), "Improvement of the excitation source in the narrow-band linear prediction vocoder", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP-33, No. 2, pp. 377–386.

A.P. Lobo and W.A. Ainsworth (1989), "Evaluation of glottal ARMA modelling scheme", *Proc. Eurospeech 89, Paris*, Vol. 2, pp. 27–30.

K. Lukaszewickz and M. Karjalainen (1987), "Microphonemic method of speech synthesis", *Proc. Int. Conf. Acoust., Speech, Signal Proc., Dallas*, p-p. 1426–1429.

J. Makhoul (1978), "A mixed source model for speech compression and synthesis", *J. Acoust. Soc. Am.* Vol. 64, No. 6, pp. 1577–1581.

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

Y. Miyoshi, K. Yamato, R. Mizogushi, M. Yanagida and O. Kakusho (1987), "Analysis of speech signals of short-time pitch period by sample selective linear prediction", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP-35, No. 9, pp. 1233–1241.

J.A. Moorer (1978), "The use of the phase vocoder in computer music applications", *J. Audio Engineer. Soc.*, Vol. JAES-26, No. 1, pp. 41–54.

E. Moulines (1990), "Algorithmes de codage et de modification des paramètres prosodiques pour la synthèse de parole à partir du texte", *Doctoral thesis, Ecole Nationale supérieure des Télécommunications*, 1990.

E. Moulines and F. Charpentier (1988), "Diphone synthesis using a multipulse LPC technique", *Proc. FASE Int. Conf., Edinburgh*, pp. 47–54.

A. Papoulis (1984), *Probability, Random Variables and Stochastic Processes* (McGraw-Hill Book Company, New York).

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

M.A. Richards (1979), "Helium speech enhancement using the short-time Fourier transform", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP-27, pp. 841–853.

S. Roucos and A. Wilgus (1985), "High quality time-scale modification for speech", *IEEE Int. Conf. Acoust., Speec, Signal Proc.*, Tampa, pp. 493–496.

Y. Sagisaka (1988), "Speech synthesis by rule using an optimal selection of non-uniform synthesis units", *IEEE Int. Conf. Acoust., Speech, Signal Proc.*, New York, pp. 679–682.

S. Seneff (1982), "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP-30, No. 4, pp. 566–577.

A. Varga and F. Fallside (1987), "A technique for using multipulse linear predictive speech synthesis in Text-To-Speech Type systems", *IEEE Trans Acoust., Speech, Signal Proc.*, Vol. ASSP-35, No. 4, pp. 586–587.