ELSEVIER

Speech Communication 18 (1996) 257–279

SPEECH COMMUNICATION

Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform [1]

Raymond Veldhuis [*], Haiyan He

Institute for Perception Research, P.O. Box 513, 5600 MB Eindhoven, Netherlands

Received 18 October 1994; revised 24 July 1995

Abstract

The modification methods described in this paper combine characteristics of PSOLA-based methods and algorithms that resynthesize speech from its short-time Fourier magnitude only. The starting point is a short-time Fourier representation of the signal. In the case of duration modification, portions, in voiced speech corresponding to pitch periods, are removed from or inserted in this representation. In the case of pitch modification, pitch periods are shortened or extended in this representation, and a number of pitch periods is inserted or removed, respectively. Since it is an important tool for both duration and pitch modification, the resynthesis-from-short-time-Fourier-magnitude-only method of Griffin and Lim (1984) and Griffin et al. (1984) is reviewed and adapted. Duration and pitch modification methods and their results are presented.

Zusammenfassung

Die hier beschriebenen Variationsmethoden kombinieren Charakteristiken von PSOLA-basierten Methoden und Algorithmen zur Resynthetisierung von Sprache allein aus ihrem Kurzzeit-Fourier-Betragsspektrum. Der Ausgangspunkt ist eine Kurzzeit-Fourier-Repräsentation des Signals. Zur Dauermodifikation werden Signalstücke ausgeschnitten, die bei stimmhafter Sprache den Pitchperioden entsprechen. Zur Pitchmodifikation werden die Pitchperioden verkürzt oder verlängert. Zum Dauerausgleich wird eine Anzahl Perioden hinzugefügt oder weggelassen. Die "resynthesis-from-short-time-Fourier-magnitude-only"-Methode von Griffin und Lim (1984) und Griffin et al. (1984) wird besprochen und angepaßt, da sie ein wichtiges Hilfsmittel sowohl zur Dauer- als auch zur Pitchveränderung ist. Methoden zur Veränderung von Dauer und Pitch und die damit erzielten Ergebnisse werden dargestellt.

Résumé

Les méthodes de modification décrites ici combinent des caractéristiques des méthodes basées sur le principe PSOLA et des algorithmes pour resynthétiser la parole basés uniquement sur le module de sa transformée de Fourier à court-terme. Le point de départ est une représentation de Fourier à court terme du signal. Pour effectuer des modifications de durée, des portions voisées correspondant à des périodes complètes sont supprimées ou insérées. En ce qui concerne les modifications de fréquence fondamentale, des périodes sont raccourcies ou ralongées et un certain nombre de périodes respectivement

[*] Corresponding author. E-mail: veldhuis@natlab.research.philips.com.

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

0167-6393/96/$15.00 Published by Elsevier Science B.V.
SSDI 0167-6393(95)00044-5

supprimées ou insérées. Comme il s'agit d'un outil important pour les modifications de durée et de fréquence fondamentale, la méthode de resynthèse de Griffin et Lim (1984) et Griffin et al. (1984), basée uniquement sur le module de la transformée de Fourier à court terme, est revue et adaptée. Des méthodes de modification de durée et de fréquence fondamentale sont présentées ainsi que leurs résultats.

Keywords: Short-time Fourier transform; Time-scale modification; Pitch modificaton; Speech processing

1. Introduction

Modifications of the duration and the pitch of speech signals are important basic tools for prosodic modification of speech, e.g. (Bailly and Benoit, 1992, Section III). An example of such a prosodic modification is the changing of intonation or duration of prerecorded carrier sentences in automatic speech-based information systems, such as traffic- or flight-information systems. Duration modification is also used in electronic dictation systems.

The first step of a modification algorithm is usually a transformation of the speech signal to another representation that lends itself better to the intended modification. This is the analysis step. After the modification the speech signal is resynthesized. Modification methods can be characterized according to the representation in which the signal is modified. Well-known representations used for duration and pitch modifications are those based on linear predictive coding (LPC) (Atal and Hanauer, 1971), pitch-synchronous overlap and add (PSOLA) methods (Hamon et al., 1989; Moulines and Charpentier, 1990) and the short-time Fourier representation (Portnoff, 1981; Griffin and Lim, 1984). We shall discuss these methods briefly.

Linear predictive coding models the speech signal as the output of a time-varying all-pole filter, excited by either sequences of equidistant pulses in the case of voiced speech, or noise in the case of unvoiced speech. In the analysis phase the filter parameters and the excitation signal are estimated repeatedly for adjacent intervals of typically 10 to 20 ms. Duration modification consists in modifying the time-scales of the filter parameters, which is done by removing or interpolating sets of them, and of the excitation signal, which is done by adding or removing portions. Pitch modification is achieved by modifying the period of the pulses in the excitation signal. Clear advantages of LPC-based methods are their simplicity compared with other methods and the fact that the LPC representation lends itself well to economic storage of speech. The latter is also a drawback: because the representation of the excitation signal as either noise or a sequence of equidistant pulses is not lossless, the quality of resynthesized speech, even without a time-scale or duration modification, deteriorates.

PSOLA-based methods decompose the speech signal as a sum of short sequences called PSOLA bells. The bells are two pitch periods long and are obtained by multiplying the signal by window functions of this length, e.g. raised cosines, centered at estimates of the excitation moments. Unlike LPC, this is a lossless representation. Duration modification is achieved by removing or repeating bells. Pitch modification is achieved by changing the distance between the centres of adjacent bells. The quality of the modified speech is fairly good, apart from some occasional roughness. PSOLA-based methods are based on the assumption that the signal is periodic. Due to this assumption, problems sometimes occur in unvoiced speech. Related but different is an overlap-add technique for duration modification based on waveform similarity (WSOLA) (Verhelst and Roelands, 1993).

The short-time Fourier transform (STFT) obtains a time-frequency representation of the speech signal. Portnoff (1981) describes a method for duration modification that is based on linear time scaling and phase modification of the short-time Fourier transform. Good results are reported at fairly large expansion (4:1) and compression (3:1) ratios. The phase modification is evaded in (Griffin and Lim, 1984; Griffin et al., 1984), where a linear time scaling is applied to the magnitude of the short-time Fourier transform. An iterative algorithm for synthesizing a signal from its short-time Fourier magnitude and a random initial phase is then used to resynthesize the speech. An extension to allow independent modifica-

tion of excitation and spectral frequency scale is presented in (Seneff, 1982).

The modification methods described in this paper combine characteristics of PSOLA-based methods (Hamon et al., 1989) and methods based on short-time Fourier transforms (Griffin and Lim, 1984; Griffin et al., 1984). The resynthesis algorithm is a variant of the algorithm for resynthesizing signals from their short-time Fourier magnitude published in (Griffin and Lim, 1984). It is adapted in such a way that it is capable of resynthesizing speech from its short-time Fourier magnitude and a partially specified phase. The starting point is a short-time Fourier representation of the signal and an estimate of the pitch period as a function of time. In the case of duration modification, portions, corresponding to pitch periods in voiced speech, are removed from or inserted in this representation. The magnitude of an inserted part is estimated on the basis of the magnitude of the short-time Fourier transform in its neighbourhood. An initial phase is computed at the position of the deletion or insertion after which the aforementioned algorithm resynthesizes the speech signal. The pitch is also modified in the short-time Fourier representation. Then the pitch periods are shortened or extended and a number of pitch periods is inserted or removed, respectively. The latter has to be done in order to keep the time-scale unchanged.

Because short-time Fourier analysis and synthesis play an important role in this paper, they will be briefly reviewed in Section 2. An iterative method for synthesis from short-time Fourier magnitude, originally presented in (Griffin and Lim, 1984), will be discussed in Section 3. Simulation results will be presented that indicate how the convergence of this method depends on the initial phase, the number of frequency points and the window shift. These results will give a better indication of the performance of this synthesis method than those presented in (Griffin and Lim, 1984; Griffin et al., 1984). As a main result it is found that, almost irrespective of the number of frequency points and the window shift chosen, this method is not very suitable for reproducing the original waveform. The resulting speech signal is intelligible but sounds noisy and rough. The quality of the reproduction improves significantly when the resynthesis algorithm is modified in such a way that part of the original phase can be specified. This will be shown in Section 4, where for various discrete-Fourier-transform sizes, window lengths and window shifts the convergence will be analysed for an artificial vowel whose phase has been specified for every other pitch period only. It has been found that if the number of frequency points is large enough, the original signal can be reproduced almost perfectly. If for every other pitch period the phase is not fully random, but is only allowed to vary randomly about its original value, almost perfect reproduction can also be obtained with shorter windows and fewer iterations. This result is important because it will later follow that shorter windows sometimes give better results. Section 5 will present a duration-modification method based on deletion or insertion of pitch periods from the signal's short-time Fourier representation. Section 6 will present a pitch-modification method that is based on extending or shortening pitch periods in the signal's short-time Fourier representation combined with deleting or adding pitch periods. The results of duration and pitch modification will be presented in Sections 5 and 6, respectively. Finally, Section 7 will present conclusions.

2. Discrete short-time Fourier analysis and synthesis

The discrete short-time Fourier transform $\{X(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ of the time signal $\{x(k)\}_{k\in\mathbb{Z}}$ is defined as follows:

$$X(m,n) = \frac{1}{\sqrt{N}} \sum_{k=-\infty}^{\infty} w_a(mS-k)\, x(k) e^{-ikn(2\pi/N)}$$

$$= e^{-imnS(2\pi/N)} \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} w_a(k) \times x(mS-k) e^{ikn(2\pi/N)},$$

$$m\in\mathbb{Z},\ n=0,\ldots,N-1. \quad (1)$$

Here $X(m,n)$ is the discrete short-time Fourier transform at time mS/f_s and at frequency $f_s n/N$, S is the window shift and f_s the sampling frequency, $\{w_a(k)\}_{k\in\mathbb{Z}}$ is a real-valued analysis window function. For the ease of notation and implementation we will omit the factors $e^{-imnS(2\pi/N)}$, use a finite-length

analysis window $w_a(k)_{k=0,\ldots,N_w-1}$ with length $N_w \leq N$ and redefine the short-time Fourier transform as

$$X(m,n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} w_a(k)\, x(mS-k) e^{ikn(2\pi/N)},$$
$$m \in \mathbb{Z}, \quad n = 0,\ldots,N-1. \tag{2}$$

It is easily recognized that $\{X(m,n)\}_{n=0,\ldots,N-1}$ is obtained via an inverse discrete Fourier transform on $\{w_a(k)\,x(mS-k)\}_{k=0,\ldots,N-1}$. The sequence $\{|X(m,n)|\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ is called the spectrogram.

It is shown in (Griffin and Lim, 1984) that the time signal can be resynthesized from its discrete short-time Fourier transform (2) by

$$x(l) = \sum_{m=-\infty}^{\infty} w_a(mS-l) \frac{1}{\sqrt{N}} \times \sum_{n=0}^{N-1} X(m,n) e^{-i(mS-l)n(2\pi/N)}, \quad l \in \mathbb{Z}. \tag{3}$$

The analysis window must satisfy

$$\sum_{m=-\infty}^{\infty} w_a^2(mS-l) = 1, \quad l \in \mathbb{Z}. \tag{4}$$

In fact, (3) in combination with (4) does not constitute a unique synthesis operator, but it can be shown that the $\{x(k)\}_{k\in\mathbb{Z}}$ obtained with (3) minimizes

$$\sum_{m=-\infty}^{\infty} \sum_{n=0}^{N-1} \left| \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} w_a(k)\, x(mS-k) e^{ikn(2\pi/N)} - X(m,n) \right|^2. \tag{5}$$

This is important when $\{X(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ is modified in such a way that it is no longer the discrete short-time Fourier transform of any time signal $\{x(k)\}_{k\in\mathbb{Z}}$.

Fig. 1(a,b) show implementations of a discrete short-time Fourier analysis and synthesis system, respectively, based on discrete Fourier transforms. The boxes D are sample-delay operators. The boxes $\downarrow S$ are decimators. Their output sample rate is a factor S lower than their input sample rate. This is achieved by only putting out every S-th sample. The boxes $\uparrow S$ increase the sample rate by a factor of S by adding $S-1$ zeros after every sample. The boxes

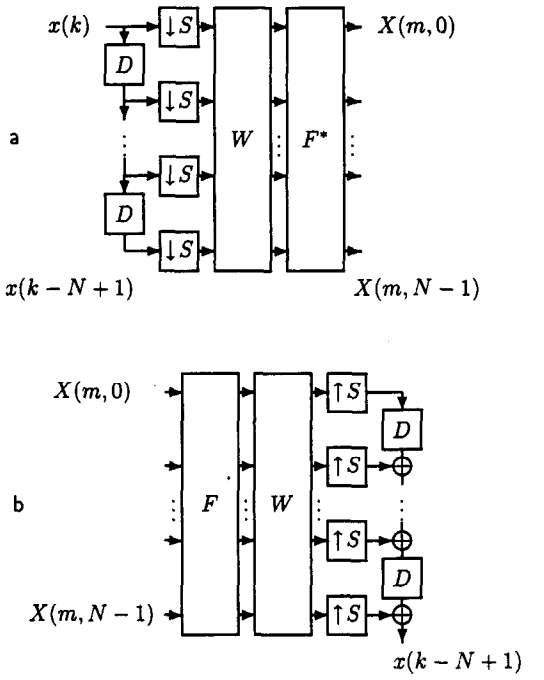


Fig. 1. (a) Discrete short-time Fourier analysis system. (b) Discrete short-time Fourier synthesis system.

W are diagonal matrices that perform the windowing. Their elements are given by

$$W_{nn} = w_a(n), \quad n = 0,\ldots,N-1. \tag{6}$$

The discrete Fourier transform and its inverse are performed by the boxes denoted F and F^*, respectively. Here F is the Fourier matrix with elements

$$F_{kl} = \frac{1}{\sqrt{N}} e^{-ikl(2\pi/N)}, \quad k,l = 0,\ldots,N-1, \tag{7}$$

and the superscript $*$ denotes complex conjugation.

3. Synthesis from discrete short-time fourier transform magnitude

The synthesis from short-time-Fourier-magnitude procedure of Griffin and Lim (1984), adapted to the discrete short-time Fourier transform pair (2) and (3), is summarized as follows. Let $\{|X_d(m,n)|\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ denote the desired spectrogram. The objective is to find a time signal

$\{x(k)\}_{k\in\mathbb{Z}}$ with a discrete short-time Fourier transform $\{X(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ such that

$$\sum_{m=-\infty}^{\infty}\sum_{n=0}^{N-1}\||X(m,n)|-|X_{\mathrm{d}}(m,n)|\|^2 \tag{8}$$

is minimum. The algorithm for obtaining $\{x(k)\}_{k\in\mathbb{Z}}$ is iterative. An initial discrete short-time Fourier transform is defined by

$$\hat{X}^{(0)}(m,n)=|X_{\mathrm{d}}(m,n)|\mathrm{e}^{\mathrm{i}\phi(m,n)},\quad m\in\mathbb{Z},\quad n=0,\ldots,N-1, \tag{9}$$

where $\phi(m,n)$ is a random phase, uniformly distributed in $[-\pi,\pi]$. In each iteration step an estimate $\{x^{(i)}(k)\}_{k\in\mathbb{Z}}$ for the time signal $\{x(k)\}_{k\in\mathbb{Z}}$ is computed from

$$x^{(i)}(k)=\sum_{m=-\infty}^{\infty}w_{\mathrm{a}}(mS-k)\frac{1}{\sqrt{N}}\times\sum_{n=0}^{N-1}\hat{X}^{(i)}(m,n)\mathrm{e}^{-\mathrm{i}(mS-k)n(2\pi/N)},\quad k\in\mathbb{Z}, \tag{10}$$

with

$$\hat{X}^{(i)}(m,n)=|X_{\mathrm{d}}(m,n)|\frac{X^{(i-1)}(m,n)}{|X^{(i-1)}(m,n)|},\quad m\in\mathbb{Z},\ n=0,\ldots,N-1, \tag{11}$$

and

$$X^{(i-1)}(m,n)=\frac{1}{\sqrt{N}}\sum_{l=0}^{N-1}w_{\mathrm{a}}(l)\,x^{(i-1)}(mS-l)\mathrm{e}^{\mathrm{i}ln(2\pi/N)},\quad m\in\mathbb{Z},\ n=0,\ldots,N-1. \tag{12}$$

The spectrogram approximation error

$$\sum_{m=-\infty}^{\infty}\sum_{n=0}^{N-1}\||X^{(i)}(m,n)|-|X_{\mathrm{d}}(m,n)|\|^2 \tag{13}$$

is a monotonically non-increasing function of i. The iterations are continued until the changes in $\{X^{(i)}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ are below a threshold. For the continuous short-time Fourier transform it is proved in (Griffin and Lim, 1984) that this iterative algorithm converges. The proof transfers directly to the discrete case.

The above method differs from the one presented in (Griffin and Lim, 1984) in a minor aspect. The method in (Griffin and Lim, 1984) starts initially with a white-noise time signal of which the short-time Fourier magnitude is modified into the desired one. The $\{\hat{X}\}^{(0)}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ that is obtained in this way has a phase $\phi(m,n)$ with dependencies in the m direction. This is a consequence of the overlap of S samples in the adjacent Fourier transforms. In the method presented here all the $\phi(m,n)$ in $\{\hat{X}\}^{(0)}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ are independent. We compared both types of initialization but found no significant differences. This is not really surprising, since after just one iteration the $\phi(m,n)$ of the present method will also show dependencies in the m direction. It will become clear from Sections 4, 5 and 6 that it is important that we specify our initial signal in the short-time Fourier domain.

A problem is that, dependent on the initial phase, the algorithm can converge to a stationary point which is not the global minimum. Starting from the spectrogram of a given speech signal the algorithm may converge to an output signal that differs significantly, in both a quadratic and a perceptual sense, from the original time signal, although the resulting spectrogram may be close to the initial one. In the following paragraphs results will be presented which show that these effects occur quite often, which makes the method unsuitable for speech synthesis or modification.

In order to be able to assess the quality of the outcome of the algorithm, we will evaluate it with a test signal $\{x_{\mathrm{d}}(k)\}_{k\in\mathbb{Z}}$ of which $\{X_{\mathrm{d}}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ is the discrete short-time Fourier transform. We define the relative mean-square error in the spectrogram after i iterations $E_{\mathrm{tf}}^{(i)}$ by

$$E_{\mathrm{tf}}^{(i)}=\frac{\sum_{m=-\infty}^{\infty}\sum_{n=0}^{N-1}\||X^{(i)}(m,n)|-|X_{\mathrm{d}}(m,n)|\|^2}{\sum_{m=-\infty}^{\infty}\sum_{n=0}^{N-1}|X_{\mathrm{d}}(m,n)|^2}, \tag{14}$$

and the relative mean-square error in the time signal after i iterations $E_{\mathrm{t}}^{(i)}$ by

$$E_{\mathrm{t}}^{(i)}=\frac{\sum_{k=-\infty}^{\infty}|x^{(i)}(k)-x_{\mathrm{d}}(k)|^2}{\sum_{k=-\infty}^{\infty}|x_{\mathrm{d}}(k)|^2}. \tag{15}$$

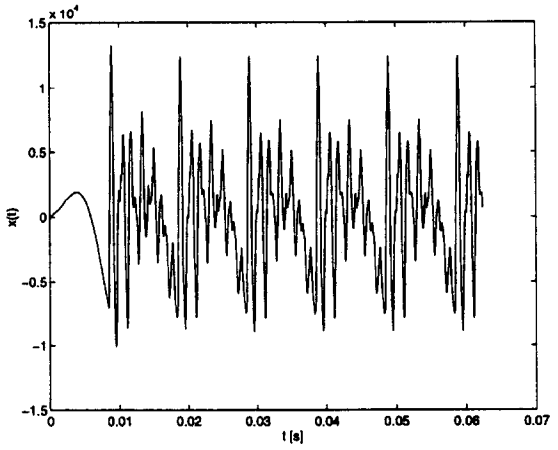


Fig. 2. Artificial vowel /a/, used as a test signal in this paper, $f_s = 16$ kHz, $f_0 = 100$ Hz.

The window that was used was the raised cosine given by

$$w_a(n) = \begin{cases} \sqrt{\dfrac{8S}{3N_w}}\,\dfrac{1-\cos\left(\frac{1}{2}(2n+1)(2\pi)/N_w\right)}{2}, \\ \qquad n = 0,\ldots,N_w-1, \\ 0, \quad n = N_w,\ldots,N-1. \end{cases} \tag{16}$$

In this manner (4) is satisfied if $S \le N_w/4$, as has also been observed in (Griffin and Lim, 1984). The parameters that were varied are the window length N_w, which was kept equal to the number of frequency points N, and the window shift S. The window length determines the trade-off between time and frequency resolution in the spectrogram. An increased window length means an increased frequency resolution and a decreased time resolution. Both N and S determine the computational complexity and the number of values generated by the short-time Fourier transform.

Both $E_{tf}^{(i)}$ and $E_t^{(i)}$ have been computed for a discrete-time signal representing an artificial vowel

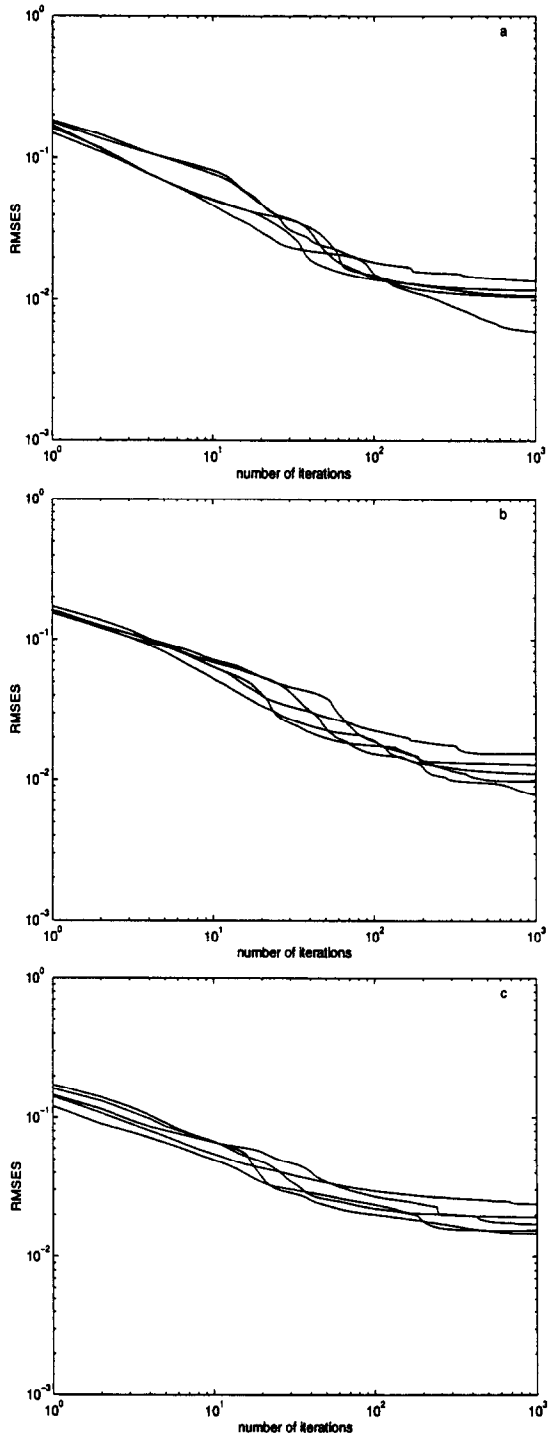


Fig. 3. Relative mean-square error in the spectrogram (RMSES) $E_{tf}^{(i)}$ as a function of i for different values chosen for the initial phase. Number of input samples 1024, window length $N_w = 128$, number of frequency points $N = 128$, window shifts (a) $S = 1$, (b) $S = 8$, (c) $S = 32$.

/a/. The sample rate f_s equals 16 kHz. The signal has a fundamental frequency $f_0 = 100$ Hz. This corresponds to a pitch period M_p of 160 samples. A part of the waveform of this signal is shown in Fig. 2.

Fig. 3 plots $E_{tf}^{(i)}$ as a function of the number of iterations i under several conditions. Fig. 4 shows $E_t^{(i)}$ as a function of the number of iterations i under the same conditions. The results in these figures were obtained using 1024 samples of the signal shown in Fig. 2. The window length N_w and the number of frequency points N were both equal to 128. Plots a, b and c in Figs. 3 and 4 show the results obtained with window shifts S equal to 1, 8 and 32, respectively. Each picture contains 5 curves. Each curve corresponds to a different pseudo random choice of the initial phase sequences $\{\phi(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$. The $\phi(m,n)$ are independent and have a uniform probability density on the interval $[-\pi,\pi]$.

Both Figs. 3 and 4 reveal the dependence on the initial phase. Most of the curves in Fig. 3, showing the relative mean-square error in the spectrogram, seem to have converged after a few hundreds of iterations. However, at lower numbers of iterations the curves show sudden jumps. It cannot be precluded that these jumps will occur again for $i > 1000$. Fig. 3 does not show a strong dependence on the window shift S, although the error is somewhat higher for $S = 32$. After 1000 iterations all relative mean-square errors in the spectrogram are close to 0.01, which corresponds to a spectral signal-to-noise ratio of 20 dB. This value has also been reported in (Griffin et al., 1984), where time-scale modification of speech was studied. It is somewhat surprising that the curves for the relative mean-square errors in the time signal, shown in Fig. 4, are not monotonic. Also, they do not seem to converge within 1000 iterations. With a relative error of about 2, it cannot be said that the output signal closely approximates the input. This is illustrated by Fig. 5, showing a

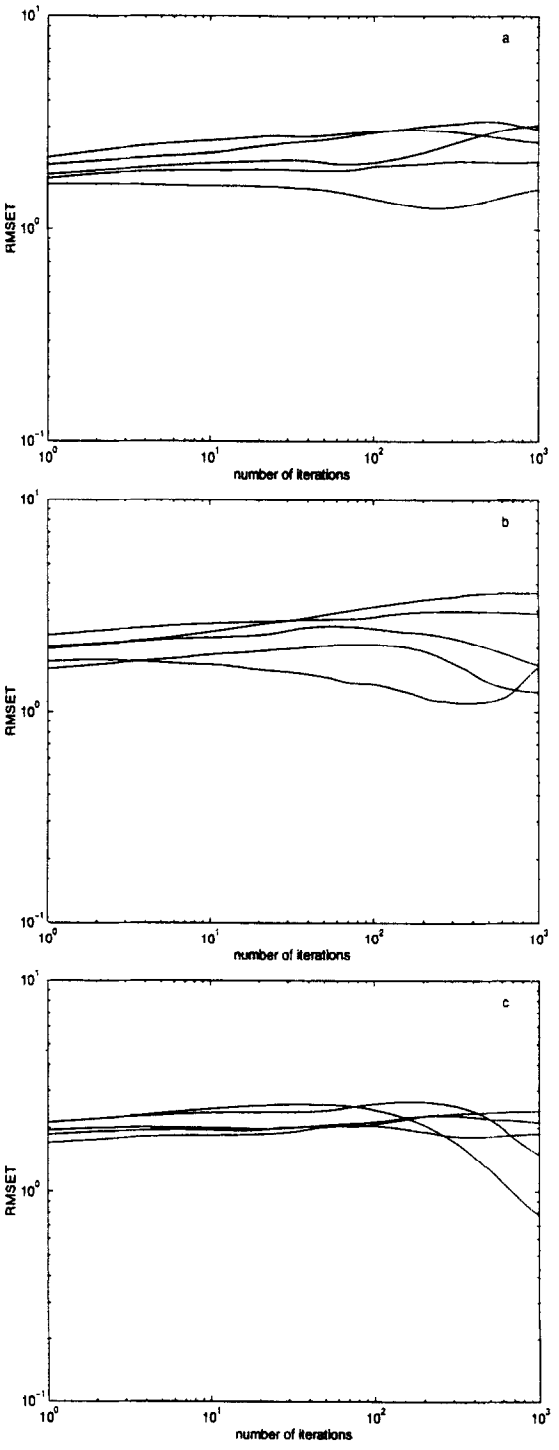


Fig. 4. Relative mean-square error in time signal (RMSET) $E_t^{(i)}$ as a function of i for different values chosen for the initial phase. Number of input samples 1024, window length $N_w = 128$, number of frequency points $N = 128$, window shifts (a) $S = 1$, (b) $S = 8$, (c) $S = 32$.

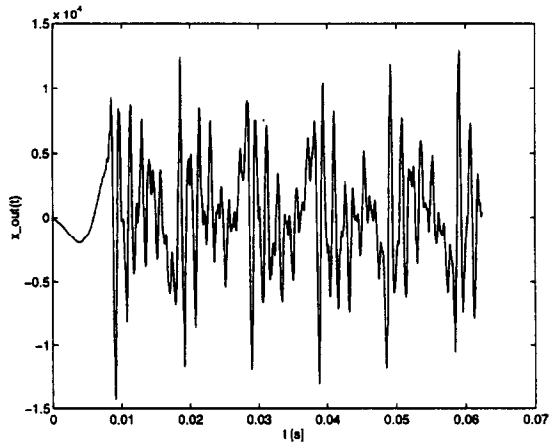


Fig. 5. Typical output for the artificial vowel /a/ obtained after 1000 iterations. Window length $N_w = 128$, number of frequency points $N = 128$, window shifts $S = 1$.

typical output signal after 1000 iterations obtained using 1024 samples of the artificial /a/, with $N_w = N = 128$, $S = 1$. The periodic structure of the signal seems to be maintained, but the waveform is not well approximated. Note the 180-degrees phase jumps that seem to change the signs of some of the pitch periods. The signal sounds like a noisy vowel /a/. This noisiness is also observed for resynthesized real speech utterances. The utterances are intelligible but of poor perceptual quality.

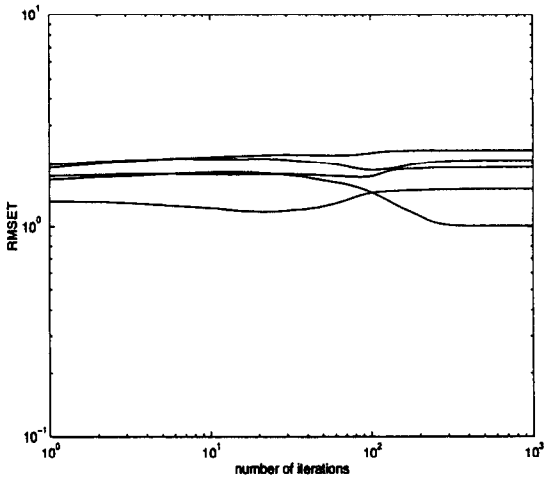


Fig. 7. Relative mean-square error in time signal (RMSET) $E_t^{(i)}$ as a function of i for different values chosen for the initial phase. Number of input samples 1024, window length $N_w = 32$, number of frequency points $N = 32$, window shift $S = 1$.

Figs. 6 and 7 show curves for $E_{tf}^{(i)}$ and $E_t^{(i)}$, respectively, obtained using the same 1024 samples of the test signal, but with $N_w = N = 32$ and $S = 1$. Although convergence seems to be reached faster than in the case of $N_w = N = 128$, the results are slightly worse.

The results shown in Figs. 8 and 9, also showing curves for $E_{tf}^{(i)}$ and $E_t^{(i)}$, respectively, but in this case

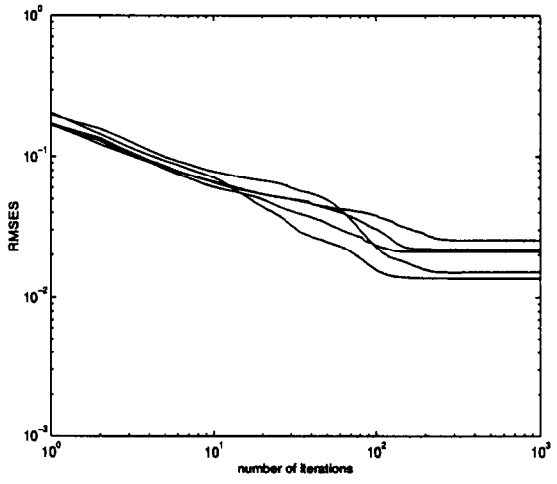


Fig. 6. Relative mean-square error in the spectrogram (RMSES) $E_{tf}^{(i)}$ as a function of i for different values chosen for the initial phase. Number of input samples 1024, window length $N_w = 32$, number of frequency points $N = 32$, window shift $S = 1$.

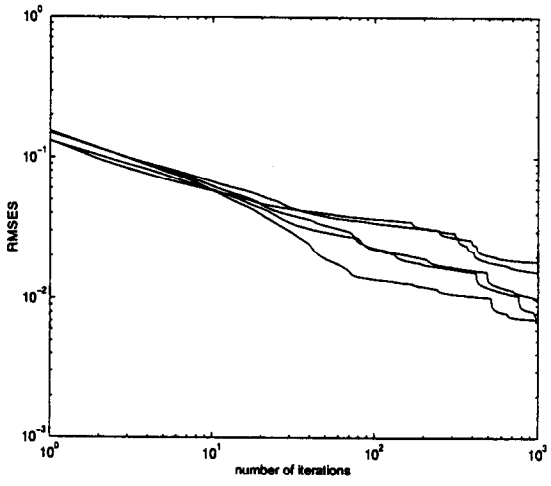


Fig. 8. Relative mean-square error in the spectrogram (RMSES) $E_{tf}^{(i)}$ as a function of i for different values chosen for the initial phase. Number of input samples 8192, window length $N_w = 1024$, number of frequency points $N = 1024$, window shift $S = 32$.

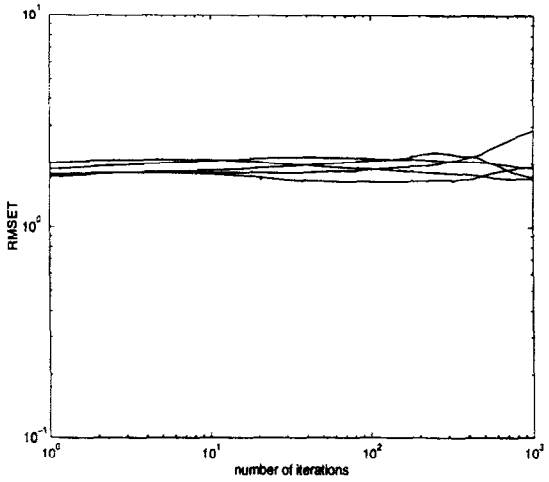


Fig. 9. Relative mean-square error in time signal (RMSET) $E_t^{(i)}$ as a function of i for different values chosen for the initial phase. Number of input samples 8192, window length $N_w = 1024$, number of frequency points $N = 1024$, window shift $S = 32$.

obtained using 8192 samples of the test signal and with $N_w = N = 1024$, $S = 32$, are also similar. It can be seen in Fig. 8 that convergence is reached more slowly when $N_w = N = 1024$ than when $N_w = N = 128$. When $N_w = N = 1024$ the curves for $E_{tf}^{(i)}$ are still decreasing at $i = 1000$. Therefore, it may be that the final approximation of the spectrogram in this case is better than for lower values of $N_w = N$. The approximation error of the input signal does not differ much between $N_w = N = 128$ and $N_w = N = 1024$.

We have also computed $E_{tf}^{(i)}$ and $E_t^{(i)}$ for $N_w = N = 256$ and $N_w = N = 512$ with $S = 1$, $S = 8$ and $S = 32$. The results were similar to those discussed above.

It seems that the number of frequency points N and the window length N_w have only a small influence on the quality of the approximated spectrogram or on the quality of the resynthesized signal, in terms of both the relative mean-square errors (14) and (15), respectively. The same seems to hold for the window shift S. The convergence speed seems to decrease with increasing $N_w = N$. It would be interesting to investigate whether an analytical relation can be obtained between convergence speed, quality of the approximations and the number of frequency points and window length.

Resynthesized sentences of real speech are of poor perceptual quality. This makes spectrogram synthesis in this form unsuitable for duration or pitch modification, as it has to compete with methods that yield substantially better quality after modifications, such as (Hamon et al., 1989; Verhelst and Roelands, 1993).

4. Synthesis with partially specified phase

The resynthesis results improve if only a part of the initial phase is random and the other part is specified correctly. The reason why this is discussed here is that it will be of importance when modification of duration and of pitch will be discussed in Sections 5 and 6, respectively. The deletion and insertion of an entire pitch period in the signal's short-time Fourier transform are basic operations in these modifications. At the location of a modification in the short-time Fourier transform the magnitude is interpolated from its neighbourhood and the phase is initially random.

The iterative procedure with a partially random initial phase is as follows. Let $\mathcal{S}$ be the set of time indices for which the initial phase is random, then the initial estimate is given by

$$\hat{X}^{(0)}(m,n) = \begin{cases} |X_d(m,n)| e^{i\phi(m,n)}, \\ \quad m \in \mathcal{S},\ n = 0,\ldots,N-1, \\ X_d(m,n), \\ \quad m \notin \mathcal{S},\ n = 0,\ldots,N-1, \end{cases} \tag{17}$$

with $\phi(m,n)$ as in (9). Iteration step (11) is replaced by

$$\hat{X}^{(i)}(m,n) = \begin{cases} |X_d(m,n)| \dfrac{X^{(i-1)}(m,n)}{|X^{(i-1)}(m,n)|}, \\ \quad m \in \mathcal{S},\ n = 0,\ldots,N-1, \\ X_d(m,n), \\ \quad m \notin \mathcal{S},\ n = 0,\ldots,N-1. \end{cases} \tag{18}$$

The same artificial vowel /a/, with a pitch period M_p of 160 samples, that was used in Section 3 has been used to compute $E_{tf}^{(i)}$ and $E_t^{(i)}$ for the synthesis algorithm with partially specified phase. The initial estimate was given by (17), the phases corresponding to every other pitch period were ran-

dom, whereas the others were copied from $\{X_d(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$. For window shifts S which are factors of M_p this corresponds to an index set $\mathcal{J}$ given by

$$\mathcal{J}=\{m|m=2aM_p/S+b,\quad a\in\mathbb{Z},\ b=0,\ldots,M_p/S-1\}. \tag{19}$$

This set corresponds to the case where every second pitch period is modified. The window was the raised-cosine window of (16). The parameters that were varied are the window length N_w, which was kept equal to the number of frequency points N, and the window shift S.

Figs. 10–13 show plots of $E_{tf}^{(i)}$ and $E_t^{(i)}$ as functions of the number of iterations under several conditions. The pictures (a), (b) and (c) show the curves obtained with window shifts given by $S=1$, $S=8$ and $S=32$, respectively. All pictures (a), (b) and (c) in Figs. 10–13 contain five curves correspondng to different pseudo-random initial phases $\{\phi(m,n)\}_{m\in\mathcal{J},n=0,\ldots,N-1}$. The $\phi(m,n)$ are independent and have uniform probability densities on the interval $[-\pi,\pi]$. Figs. 10 and 11 were obtained with the window length N_w and the number of frequency points N given by $N_w=N=128$. The number of samples in the test signal was 1024. Figs. 12 and 13 were obtained with the window length N_w and the number of frequency points N given by $N_w=N=256$. The number of samples in the test signal was 2048.

Compared with Fig. 3, Fig. 10 shows a somewhat better approximation of the spectrogram. Dependent on the initial phase, the spectrogram approximation is sometimes much better, e.g. Fig. 10(a), lower curve. Fig. 10 does not show a strong dependence on the window shift S, although for $S=32$ the performance seems to deteriorate slightly. The approximation of the time signal shown in Fig. 11 has improved relative to the approximation with completely random phase, as shown in Fig. 4. For all window

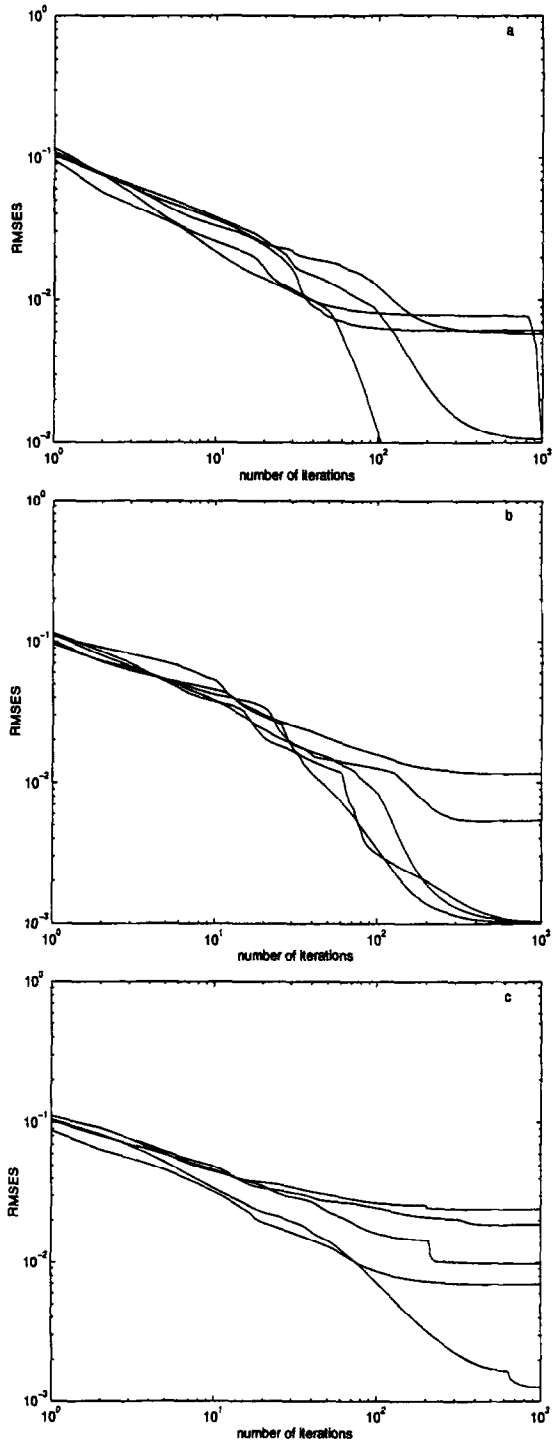
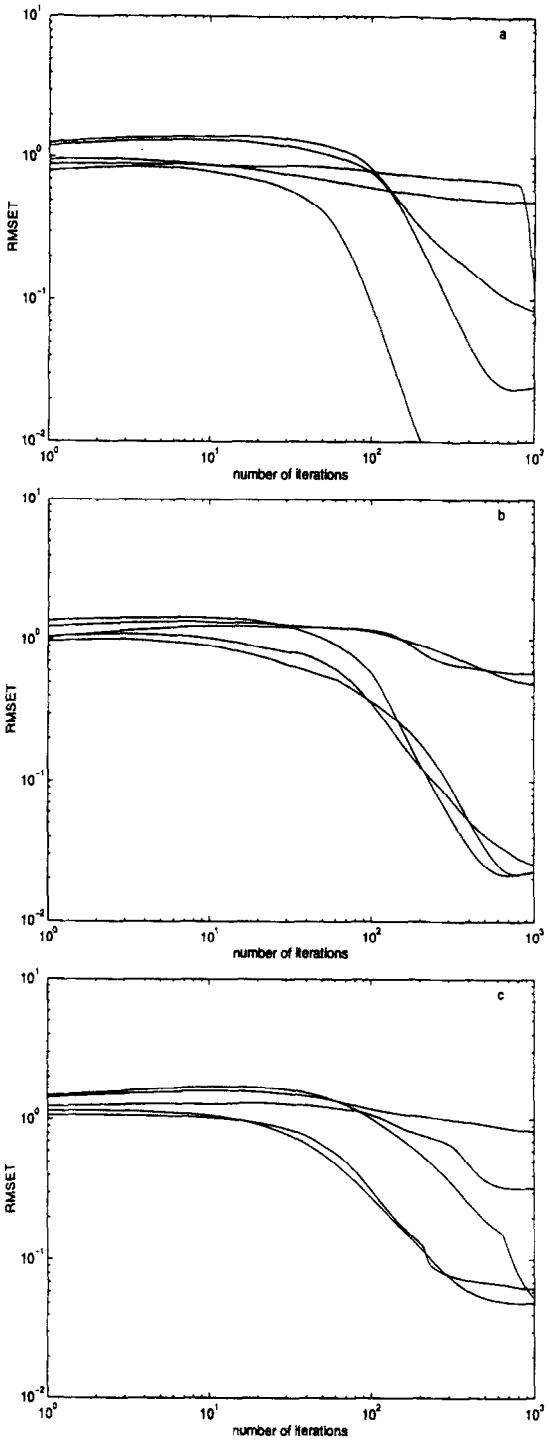


Fig. 10. Relative mean-square error in the spectrogram (RMSES) $E_{tf}^{(i)}$ as a function of i for different values chosen for the initial phase. The initial phase was random for every other pitch period. Number of input samples 1024, window length $N_w=128$, number of frequency points $N=128$, window shifts (a) $S=1$, (b) $S=8$, (c) $S=32$.



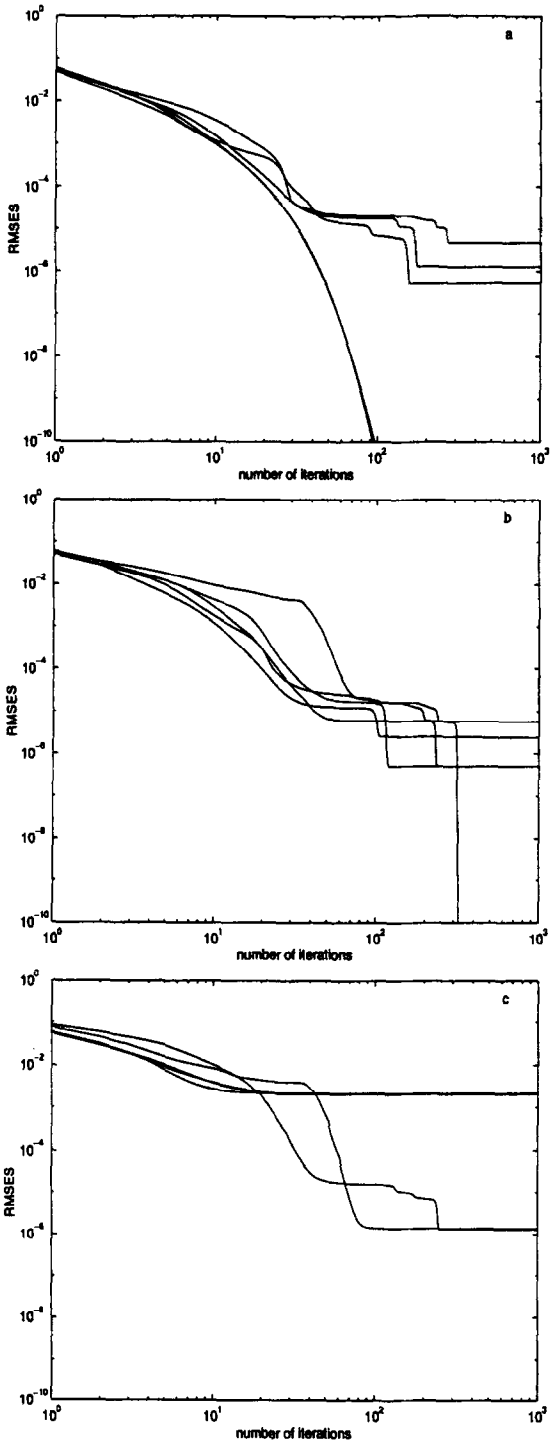
shifts $E_t^{(i)}$ is between 0.02 and 0.8, with the exception of the lower curve in Fig. 11(a). The improvement is by a factor of between 2 and 50. However, in terms of signal-to-noise ratios, the quality of the resynthesized signal is still not better than 20 dB, which is not enough for high-quality resynthesis. The performance improves further if the window length and the number of frequency points are increased to $N_w = N = 256$. Figs. 12 and 13 illustrate the improvement. For $S = 1$ and $S = 8$, the spectrogram errors are about 10^{-5} and the errors in the time signal are about 10^{-4}, with some positive exceptions, where both are computed as 10^{-28}. With due allowance for computational rounding errors, this means a perfect reconstruction. For $S = 32$ the errors are bigger, again with some exceptions. For $S = 1$ and $S = 8$, the signal-to-noise ratios of the resynthesized time signals are about 40 dB, which is quite acceptable. There is still some dependence on the initial phase, but it is less important, since the worst approximations are good enough. Convergence seems to be reached after about 200 iterations. It can be concluded from the lower curves in Fig. 12(a,b) and Fig. 13(a,b) that there are some very fortunate choices of values for the initial phase.

The reason why the choice $N_w = N = 256$ leads to so much better performance than $N_w = N = 128$ is unclear. A possible reason is that we considered a signal with a pitch period containing $M_p = 160$ samples. For $N_w = N = 256$ we have $N_w = N > M_p$, but for $N_w = N = 128$ we have $N_w = N < M_p$. If we regard the analysis/synthesis system from a filterbank point of view, we can derive that the $\{X(m,n)\}_{m \in Z, n = 0, \ldots, N-1}$ can be written as

$$X(m,n) = \sum_{k=-\infty}^{\infty} h_n(mS - k)\, x(k), \quad m \in Z,$$

$$n = 0, \ldots, N-1, \tag{20}$$

Fig. 11. Relative mean-square error in time signal (RMSET) $E_t^{(i)}$ as a function of i for different values chosen for the initial phase. The initial phase was random for every other pitch period. Number of input samples 1024, window length $N_w = 128$, number of frequency points $N = 128$, window shifts (a) $S = 1$, (b) $S = 8$, (c) $S = 32$.



with the analysis filters given by

$$h_n = w_a(k)e^{ikn(2\pi/N)}, \quad n = 0,\ldots,N-1, \quad k = 0,\ldots,N-1. \tag{21}$$

Generally speaking, if $S < N_w = N$, the $\{X(m,n)\}_{m\in Z, n=0,\ldots,N-1}$ are redundant in the time direction. Therefore, information on the phase in the unspecified parts is contained in the specified parts. The resynthesized signal can be written as

$$x(l) = \sum_{n=0}^{N-1} \sum_{m=-\infty}^{\infty} g_n(l - mS)\, X(m,n), \quad l \in \mathbb{Z}, \tag{22}$$

with the synthesis filters given by

$$g_n(k) = w_a(N-1-k)e^{-i\,(N-1)-k)n(2\pi/N)}, \quad n = 0,\ldots,N-1,\ k = 0,\ldots,N-1. \tag{23}$$

This means that if $N_w = N > M_p$, then the synthesis filters are more capable of copying correct phase information to the unspecified parts. Of course, this is an assumption that needs further investigation.

The following results are also of importance for the next sections. The relatively large number of frequency points $N = 256$, combined with a window shift $S = 1$ and a number of iterations that is greater than 200 imply a long computation time, even on fast processors. For practical applications such as the pitch and duration modification, that have to run in, or close to, real time, this is a problem. It will therefore be investigated whether an initial phase that is already close to the true phase, combined with a smaller number of frequency points N will lead to acceptable results with less computational effort. The situation of an initial phase close to the true one is realistic in the cases of pitch and duration modification, because if the signal is periodic, a good estimate for the initial phase at the location of a modification can be obtained via interpolation of the short-time Fourier transform.

Fig. 12. Relative mean-square error in the spectrogram (RMSES) $E_{tf}^{(i)}$ as a function of i for different values chosen for the initial phase. The initial phase was random for every other pitch period. Number of input samples 2048, window length $N_w = 256$, number of frequency points $N = 256$, window shifts (a) $S = 1$, (b) $S = 8$, (c) $S = 32$.

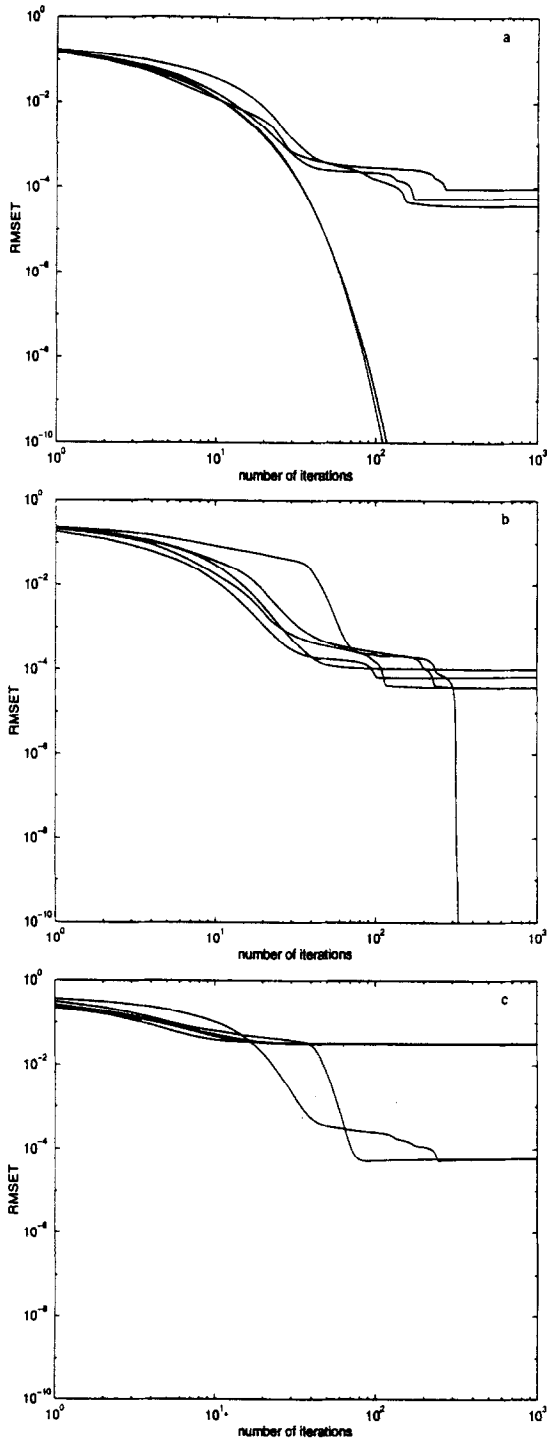


Fig. 13. Relative mean-square error in time signal (RMSET) $E_t^{(i)}$ as a function of i for different values chosen for the initial phase. The initial phase was random for every other pitch period. Number of input samples 2048, window length $N_w = 256$, number of frequency points $N = 256$, window shifts (a) $S = 1$, (b) $S = 8$, (c) $S = 32$.

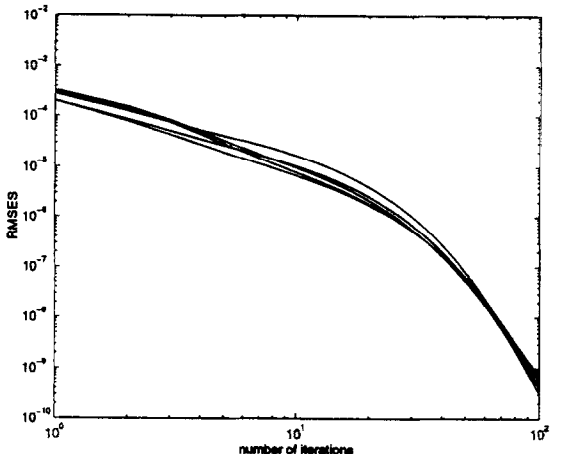


Fig. 14. Relative mean-square error in the spectrogram (RMSES) $E_{tf}^{(i)}$ as a function of i for different values chosen for the initial phase. For every other pitch period the initial phase has a small random error. Number of input samples 1024, window length $N_w = 32$, number of frequency points $N = 32$, window shift $S = 1$.

Figs. 14 and 15 show curves for $E_{tf}^{(i)}$ and $E_t^{(i)}$, respectively, obtained using the same 1024 samples of the test signal, but with $N_w = N = 32$ and $S = 1$. The window is the raised cosine window of (16). The algorithm is the one used for synthesis with partially random phase that has been described earlier in this section. The difference concerns the initial estimate for the phase, which is now the original phase with a small random component added to it. This means that (17) has been replaced by

$$\hat{X}^{(0)}(m,n) = \begin{cases} |X_d(m,n)|e^{i(\arg(X_d(m,n))+\phi(m,n))}, \\ \quad m \in \mathcal{J},\ n = 0,\ldots,N-1, \\ X_d(m,n), \\ \quad m \notin \mathcal{J},\ n = 0,\ldots,N-1, \end{cases} \tag{24}$$

with $\mathcal{J}$ given by (19) and the $\phi(m,n)$ independent random variables, uniformly distributed in

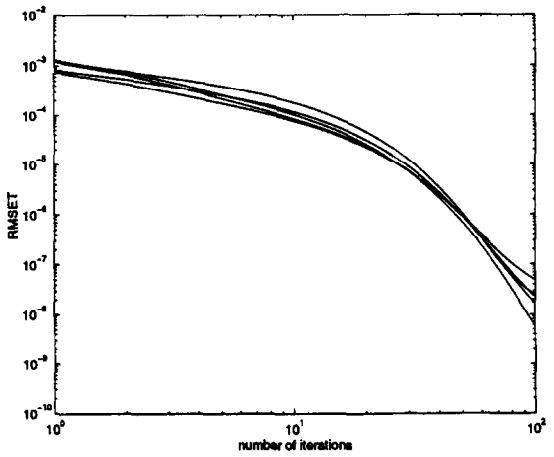


Fig. 15. Relative mean-square error in time signal (RMSET) $E_t^{(i)}$ as a function of i for different values chosen for the initial phase. For every other pitch period the initial phase has a small random error. Number of input samples 128, window length $N_w = 32$, number of frequency points $N = 32$, window shift $S = 1$.

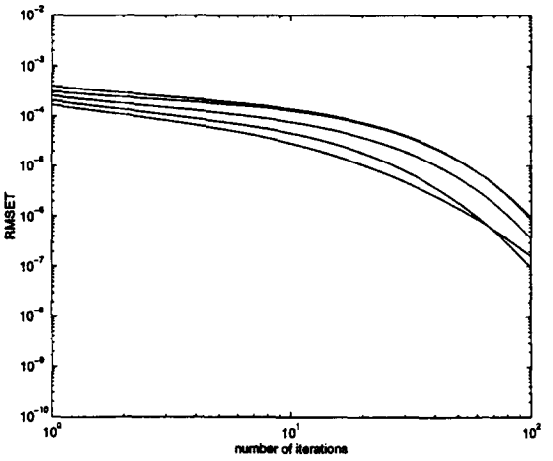


Fig. 17. Relative mean-square error in time signal (RMSET) $E_t^{(i)}$ as a function of i for different values chosen for the initial phase. For every other pitch period the initial phase has a small random error. Number of input samples 1024, window length $N_w = 32$, number of frequency points $N = 128$, window shift $S = 1$.

$[-\alpha\pi, \alpha\pi]$. The phase error is controlled by α. An α close to zero means an initial estimate for the phase close to the original, an α close to one brings us back to the situation described earlier in this section.

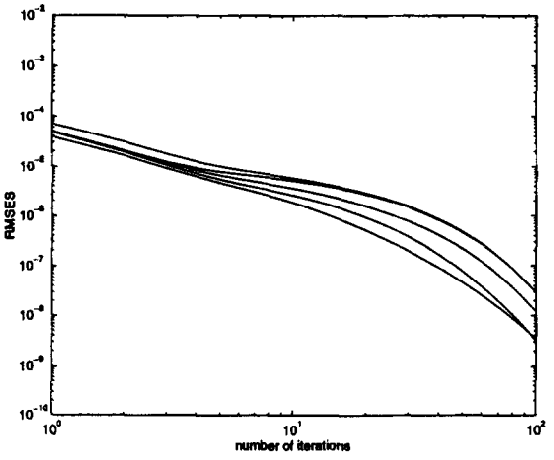


Fig. 16. Relative mean-square error in the spectrogram (RMSES) $E_{tf}^{(i)}$ as a function of i for different values chosen for the initial phase. For every other pitch period the initial phase has a small random error. Number of input samples 1024, window length $N_w = 32$, number of frequency points $N = 128$, window shift $S = 1$.

In Figs. 14 and 15 we have $\alpha = 0.2$. Figs. 16 and 17 also show curves for $E_{tf}^{(i)}$ and $E_t^{(i)}$. The difference with Figs. 14 and 15 is that the curves were obtained with analysis and synthesis windows of a length N_w which is less than the number of frequency points N. In Figs. 16 and 17 we have $N = 128$, but $N_w = 32$. The other conditions are identical.

The curves in Figs. 14–17 show faster convergence rates, for both the approximation of the spectrogram and the time signal, than those obtained with a partially random phase. This was to be expected, because the initial phase estimate is closer to the original. Comparing the curves of Figs. 14–17 with those in Fig. 12(a) and Fig. 13(a), obtained with longer analysis and synthesis windows ($N = 256$), we observe that we reach approximately the same error level in spectrogram and time signal, but after 5–10 iterations instead of after about 50. Together with the shorter windows this implies a large reduction in computational effort. The curves in Figs. 16 and 17, which were obtained with a number of frequency points $N = 128$ and a window size $N_w = 32$, show faster initial convergence rates than those in Figs. 14 and 15 where $N = N_w = 32$. Error levels $E_{tf}^{(i)} = 10^{-5}$ and $E_t^{(i)} = 10^{-4}$ are reached about twice as fast. Surprisingly, after more iterations their convergence rate becomes slower.

5. Duration modification

In PSOLA-based duration-modification methods (Hamon et al., 1989), deleting pitch periods from and inserting them into the time signal are the basic operations that are repeated according to a certain strategy. In the case of an insertion, the inserted pitch period is usually a copy of a pitch period in the neighbourhood. Here we will present a duration-modification method that deletes pitch periods from or inserts them into the short-time Fourier transform. This is done in such a way that the short-time-Fourier-transform magnitude is specified everywhere and a good approximate initial phase is chosen at the position of the deletion and the insertion. This is the situation where we have a partially specified initial phase with the unspecified parts being a good approximation of the original phase. This situation is similar to the one that led to the synthesis algorithm of Section 4, with (24) specifying the initial phase. This algorithm will be the basis for our duration manipulation method.

The basic deletion and insertion operations will be described first. It is required that a reliable estimate of the pitch period be available as a function of time. Let this estimate be denoted by $\{M_p(m)\}_{m\in\mathbb{Z}}$. If confusion is not likely to arise, we will use just M_p for the local pitch period. The pitch-estimation method should be such that in unvoiced intervals an estimate, which may be arbitrary, is available too, e.g. the one described in (Hermes, 1988). In addition a voiced/unvoiced indication is required. The original short-time Fourier transform is denoted by $\{X_{\mathrm{org}}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$. Everywhere we have $S = 1$, so that an index set $\mathcal{J}$ according to (19) can always be found.

Let us assume first that we want to delete $\{X(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ over the length of M_p samples starting at time index m_0. We define as an initial estimate

$$\hat{X}^{(0)}(m,n) = \begin{cases} X_{\mathrm{org}}(m,n), \\ \quad m < m_0,\ n = 0,\ldots,N-1, \\ X_{\mathrm{org}}(m+M_p,n), \\ \quad m \geq m_0,\ n = 0,\ldots,N-1, \end{cases} \tag{25}$$

choose

$$\mathcal{J} = \{m \mid m_0 - m_p < m \leq m_0 + M_p\}, \tag{26}$$

and repeat iteration steps (10), (18) and (12). The index set $\mathcal{J}$ refers to the time indices of the $\{X^{(i)}(m,n)\}_{i\geq 0,m\in\mathbb{Z},n=0,\ldots,N-1}$ and $\{\hat{X}^{(i)}(m,n)\}_{i\geq 0,m\in\mathbb{Z},n=0,\ldots,N-1}$. The value chosen for $\mathcal{J}$ is rather arbitrary. A somewhat larger or smaller index set also satisfies. Note that the iteration changes the time signal over the interval $[m_0 - M_p - N/2, m_0 + M_p + N/2]$. This interval will be called the modified interval.

The following procedure is used to insert a pitch period at time index m_0 in voiced speech. The initial estimate is given by

$$\hat{X}^{(0)}(m,n) = \begin{cases} X_{\mathrm{org}}(m,n), \\ \quad m < m_0,\ n = 0,\ldots,N-1, \\ \frac{1}{2}\big(|X_{\mathrm{org}}(m-M_p,n)| + |X_{\mathrm{org}}(m,n)|\big)e^{i\phi(m,n)}, \\ \quad m_0 \leq m < m_0 + M_p,\ n = 0,\ldots,N-1, \\ X_{\mathrm{org}}(m-M_p,n), \\ \quad m \geq m_0 + M_p,\ n = 0,\ldots,N-1. \end{cases} \tag{27}$$

For the initial phase we choose

$$\phi(m,n) = \arg\big(X_{\mathrm{org}}(m-M_p,n) + X_{\mathrm{org}}(m,n)\big), \quad m_0 \leq m < m_0 + M_p,\ n = 0,\ldots,N-1. \tag{28}$$

These initial estimates are good if $\{X_{\mathrm{org}}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ is quasi-periodic in m with period M_p. In unvoiced speech we choose as an initial estimate

$$\hat{X}^{(0)}(m,n) = \begin{cases} X_{\mathrm{org}}(m,n), \\ \quad m < m_0, \\ \big((1-\gamma)|X_{\mathrm{org}}(m_0-1,n)| \\ \quad + \gamma|X_{\mathrm{org}}(m_0,n)|\big)e^{i\phi(m,n)}, \\ \quad m_0 \leq m < m_0 + M_p, \\ X_{\mathrm{org}}(m-M_p,n), \\ \quad m \geq m_0 + M_p, \end{cases} \tag{29}$$

with $n = 0, \ldots, N-1$ and

$$\gamma = \frac{m - m_0 + 1}{M_p}. \tag{30}$$

The initial phase $\phi(m,n)$ is random, as in (9). The linear interpolation in the initial estimate aims to realize a smooth spectrogram. In both the voiced and unvoiced case the index set $\mathcal{S}$ is given by

$$\mathcal{S} = \{m | m_0 \le m < m_0 + M_p\}. \tag{31}$$

The iteration steps (10), (18) and (12) are repeated. The modified interval is given by $[m_0 - N/2, m_0 + M_p + N/2]$.

Unlike in PSOLA (Hamon et al., 1989), neither insertion nor deletion of pitch periods requires an estimate of the excitation moment. However, inserting or deleting pitch periods at positions where the spectrogram changes much as a function of time may cause audible effects. Therefore insertion or deletion points are placed at positions within a pitch period where the spectral change in the time direction is small. A spectral change measure that can be used to determine such a point is

$$D_{tf}(m) = \sum_{n=0}^{N-1} \| |X(m,n)| - |X(m-1,n)| \|, \quad m \in Z. \tag{32}$$

The next paragraphs will present some results obtained with the duration modification method described before. The position within a pitch period with the minimum spectral change $D_{tf}(m)$ defined by (32) was taken for the point of a deletion or insertion. The pitch was estimated with the aid of an improved version of the method described in (Hermes, 1988), which also provides a voiced/unvoiced indication. The results can only be good if the distance between two insertion or deletion points is larger than N. This means that the duration modification was performed in steps, in each of which the modified intervals did not overlap.

Fig. 18 shows 1000 samples of the artificial vowel /a/ of Fig. 2 that has been extended by a factor of two. The extension was obtained by inserting one pitch period after every original pitch period. The window was a raised cosine, given by (16), with $N_w = 32$. The number of frequency points was given by $N = 128$. The number of iterations was 5. From the figure it cannot be seen which pitch periods have been inserted. Informal listening does not reveal audible differences between the original vowel and the extended one.

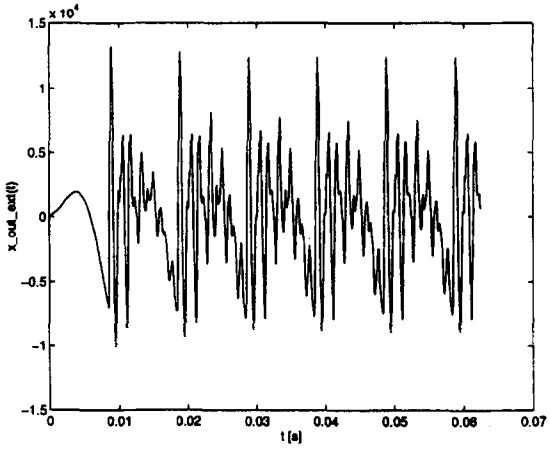


Fig. 18. Artificial vowel /a/ with the duration extended to 200%. Window length $N_w = 32$, number of frequency points $N = 128$, window shifts $S = 1$, number of iterations 5.

Figs. 19–21 show an original, a 50%-shortened and a 100%-extended version of the Dutch word 'toch', /t ⊃ χ/, pronounced by a male voice, respectively. The sample rate was 10 kHz, instead of 16 kHz for the artificial vowel. The window was a raised cosine, given by (16), with $N_w = 64$. The number of frequency points was given by $N = 512$. The number of iterations was 30. The results presented in Figs. 16 and 17 suggest that a smaller number may have been sufficient. Indeed, most of the time fewer than 5 iterations suffice. Only if the pitch estimate is incorrect, which happens rarely, more iterations do seem to improve the result.

The quality was judged in informal listening tests only. In these tests, the time-scale of Dutch and English sentences, uttered by various male as well as female voices, was changed between a reduction to 20% and an extension to 300% of the original length [2]. A comparison was made with a time-do-

[2] Examples of sentences used in this test are accessible from the Elsevier WWW server (http://www.elsevier.nl/locate/specome). Signal A is the original sentence "The goose laid an odd egg" uttered by a male speaker at a conversational rate; signal B results from a time-scale reduction to 66%; signal C results from a time-scale extension to 150%.

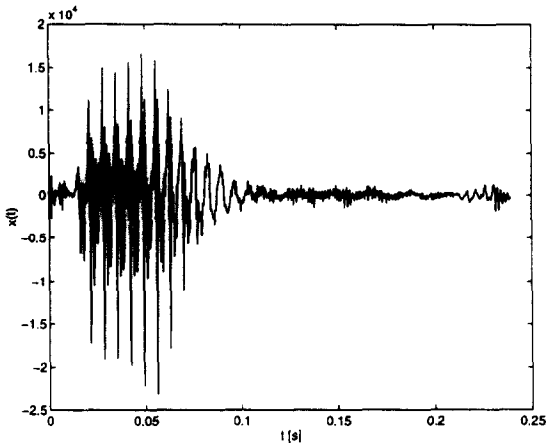


Fig. 19. Original version of the Dutch word 'toch', /t ɔ χ/, $f_s = 10$ kHz.

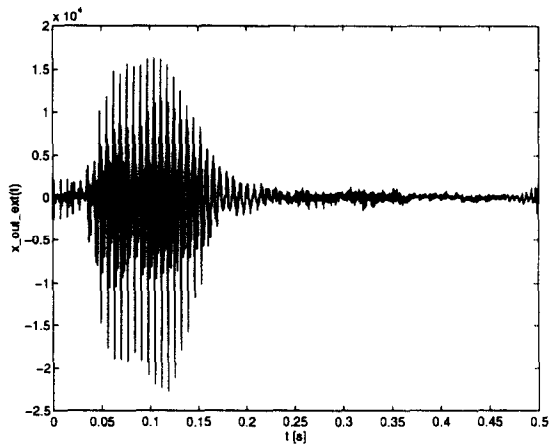


Fig. 21. The Dutch word 'toch', /t ɔ χ/, with the duration extended to 200%. Window length $N_w = 64$, number of frequency points $N = 512$, window shifts $S = 1$, number of iterations 30.

main PSOLA method (TD-PSOLA). The TD-PSOLA method used Hanning windows that were precisely two pitch periods long and were centered around estimated moments of excitation. A portion of the thus obtained short-term signals was either repeated or deleted, dependent on whether the signal was extended or shortened. The evaluation was done in a quiet environment by experienced listeners. Between a reduction to 50% and an extension to 200% the quality was considered to be good. Outside this range some deteriorations became audible. In all cases the results were preferred to those obtained with the TD-PSOLA method. Especially when the time-scale was modified more than 50% in either direction, the TD-PSOLA method produced a certain roughness in vowels and some deteriorations in unvoiced sounds and voiced fricatives that were not perceived with the present duration-modification method. The results seem to be somewhat dependent on the choice of the number of frequency points N and the window length N_w chosen. The number of frequency points, $N = 512$, can be reduced to 128 at the expense of some slight deteriorations in unvoiced fricatives. The performance for female voices improves if we take $N_w = 32$, rather than $N_w = 64$.

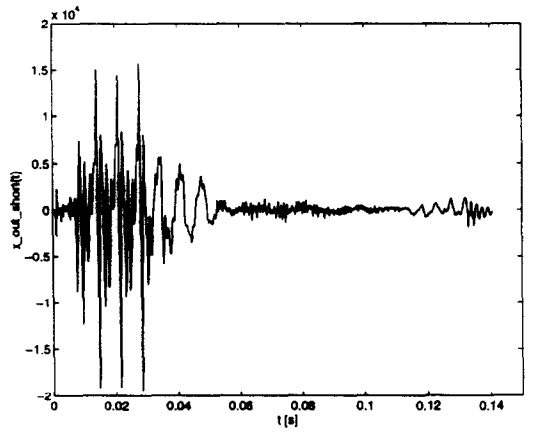


Fig. 20. The Dutch word 'toch', /t ɔ χ/, with the duration reduced to 50%. Window length $N_w = 64$, number of frequency points $N = 512$, window shifts $S = 1$, number of iterations 30.

We have also done some evaluations with interfering white noise and interfering speech. The parameter settings were identical to those for clean speech. The pitch of the corrupted signal was estimated. In both cases the interference level was −14 dB. The method seems very robust for these types of interferences, and, especially in the case of interfering speech, clearly outperforms TD-PSOLA.

6. Pitch modification

Pitch modification in the short-time Fourier representation is a two-step procedure. One step consists of shortening or extending pitch periods. This step will be discussed in this section. The other step,

which consists of inserting or deleting entire pitch periods, has been discussed in Section 5. When the pitch is decreased by a fraction, the first step is to reduce the number of pitch periods by this fraction and the second to increase the length of each pitch period by the same fraction. When the pitch is increased by a fraction, the first step is to decrease the length of each pitch period by this fraction and the second is to increase the number of pitch periods by the same fraction.

As in Section 5, it is required that a reliable estimate of the pitch period as a function of time $\{M_p(m)\}_{m\in\mathbb{Z}}$ be available. The desired pitch period is $\{M'_p(m)\}_{m\in\mathbb{Z}}$. The pitch-estimation method that is used should be such that an estimate, which may be arbitrary, is available in unvoiced intervals too, e.g. the one described in (Hermes, 1988). A voiced/unvoiced indication is also required. The original short-time Fourier transform is denoted by $\{X_{org}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$. We have $S=1$ everywhere.

When increasing the pitch we denote the number of time indices by which the pitch periods in the original short-time Fourier transform $\{X_{org}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ are to be reduced by

$$\Delta_p^-(m) = M_p(m) - M'_p(m),\quad m\in\mathbb{Z}. \tag{33}$$

When decreasing the pitch we denote the number of time indices by which the pitch periods in the original short-time Fourier transform $\{X_{org}(m,n)\}_{m\in\mathbb{Z},n=0,\ldots,N-1}$ are to be extended by

$$\Delta_p^+(m) = M'_p(m) - M_p(m),\quad m\in\mathbb{Z}. \tag{34}$$

Finding the points in the short-time Fourier transform at which the pitch period can be reduced or extended is a problem. This problem arises particularly in the case of voiced speech. In the case of unvoiced speech the points of insertion or deletion are not critical. In the case of an insertion, finding the values with which the short-time Fourier transform must be extended is an additional problem. We will use a source-filter model for speech to tackle these problems. In this source-filter model, speech is considered to be the output of a time-varying all-pole filter, modelling the vocal tract, followed by a differentiator modelling the radiation at the lips. This system is excited by a quasi-periodic sequence of glottal pulses in the case of voiced speech or by noise in the case of unvoiced speech, cf. e.g. (O'Shaughnessy, 1990, p. 80). In the open phase of a glottal cycle air flows through the glottis. In the closed phase the speech signal is solely determined by the properties of the vocal tract. This suggests that the best points for removing a portion from or inserting a portion in the pitch period are at the end of the closed phase, just before the following glottal pulse starts to influence the speech signal. We will try to determine these points in the short-time Fourier transform. Therefore, the pitch must be resolved in the time direction, which means that the window length N_w must be shorter than a pitch period. In fact, it is necessary to require that pitch be unresolved in frequency direction, otherwise the resynthesized signal will retain the old pitch.

The approach that we will use is fairly pragmatic. We will assume that the window length is shorter than the closed phase of the glottal cycle. Then, during the closed phase, the spectrogram will not contain sharp transitions. This means that $D_{tf}(m)$, defined in (32), will be small. We will measure a total $D_{tf}(m)$ over an interval to determine the points for removing or inserting portions. Even if the assumptions are not justified, it seems a fairly safe approach to modify the short-time Fourier transform in those regions where changes in the temporal direction are small.

Let us assume, for the ease of notation, that we only want to shorten or extend one pitch period at time index m_0. If we shorten a pitch period we choose m_0 as the value of m that minimizes

$$V_{tf}^-(m) = \sum_{k=m}^{m+\Delta_p^-(m)-1} D_{tf}(k), \tag{35}$$

over a pitch period. This implies that m_0 is at the beginning of a portion of the short-time Fourier transform with little variation in temporal direction. We define as an initial estimate

$$\hat{X}^{(0)}(m,n) = \begin{cases} X_{org}(m,n), \\ \quad m < m_0,\ n=0,\ldots,N-1, \\ X_{org}\big(m+\Delta_p^-(m_0),n\big), \\ \quad m \geq m_0,\ n=0,\ldots,N-1, \end{cases} \tag{36}$$

choose

$$\mathcal{J} = \mathbb{Z}, \tag{37}$$

and repeat iteration steps (10), (18) and (12). The index set $\mathcal{J}$ refers to the time indices of the $\{X^{(i)}(m,n)\}_{i \geq 0, m \in \mathbb{Z}, n = 0, \ldots, N-1}$ and $\{\hat{X}\}^{(i)}(m,n)\}_{i \geq 0, m \in \mathbb{Z}, n = 0, \ldots, N-1}$. Note that we allow the phase to change everywhere during the iterations. This is the easiest solution, since in this case we cannot use an $\mathcal{J}$ such as (26). No distinction is made between voiced and unvoiced speech.

If we extend a pitch period, we choose m_0 as the value of m that minimizes

$$V_{\mathrm{tf}}^{+}(m) = \sum_{k = m - \lfloor \beta M_{\mathrm{p}}(m) \rfloor}^{m-1} D_{\mathrm{tf}}(k), \tag{38}$$

over a pitch period. Here β is a fixed estimate of the fraction of the glottal cycle that is closed. We have taken $\beta = 1/3$. This implies that m_0 is at the end of a portion of the short-time Fourier transform with little variation in temporal direction. In this case there is the additional problem of computing the initial estimate

$$\left\{\hat{X}(m,n)\right\}_{m = m_0, \ldots, m_0 + \Delta_{\mathrm{p}}^{+}(m_0) - 1, n = 0, \ldots, N-1}. \tag{39}$$

We will make a distinction between voiced and unvoiced speech. Ideally, for voiced speech during a relaxation period the speech sample $x(k)$ is given by

$$x(k) = \sum_{l=1}^{p} a_l x(k-l), \tag{40}$$

with p being the order of the all-pole filter and the $\{a_l\}_{l=1,\ldots,p}$ the prediction coefficients. For real-valued signals we have $a_l \in \mathbb{R}$, $l = 1, \ldots, p$. We will assume a similar predictive model for the short-time Fourier transform in the relaxation period:

$$\begin{aligned} X(m,n) &= \sum_{l=1}^{p_n} a_{n,l} X(m-l,n), \\ m &= m_0 - \left\lfloor \beta M_{\mathrm{p}}(m_0) \right\rfloor, \ldots, m_0 - 1, \\ n &= 0, \ldots, N-1, \end{aligned} \tag{41}$$

with $a_{n,l} \in \mathbb{C}$, $n = 0, \ldots, N-1$, $l = 1, \ldots, p_n$, and will use (41) to extend $\{X(m,n)\}_{n=0,\ldots,N-1}$ for $m \geq m_0$. The choice $p_n = 4$, $n = 0, \ldots, N-1$, has proved to yield acceptable results. The complex prediction coefficients are estimated from

$$\left\{X(m,n)\right\}_{m = m_0 - \lfloor \beta M_{\mathrm{p}}(m_0) \rfloor, \ldots, m_0 - 1, n = 0, \ldots N-1}, \tag{42}$$

by means of Burg's method (Marple, Jr., 1987). For voiced speech we define as an initial estimate

$$\hat{X}^{(0)}(m,n) = \begin{cases} X_{\mathrm{org}}(m,n), \\ \quad m < m_0, \ n = 0, \ldots, N-1, \\ \sum_{l=1}^{p_n} a_{n,l} \hat{X}^{(0)}(m-l,n), \\ \quad m_0 \leq m < m_0 + \Delta_{\mathrm{p}}^{+}(m_0), \\ \quad n = 0, \ldots, N-1, \\ X_{\mathrm{org}}\left(m - \Delta_{\mathrm{p}}^{+}(m_0), n\right), \\ \quad m \geq m_0 + \Delta_{\mathrm{p}}^{+}(m_0), \ n = 0, \ldots, N-1. \end{cases} \tag{43}$$

We become less dependent on the point of the insertion, which has to be at the end of the relaxation period, if we use an interpolation method, e.g. (Janssen et al., 1986), instead of an extrapolation method in (43). This could be a useful improvement. In the unvoiced case the initial estimate is given by (29) and (30), with M_{p} being replaced by $\Delta_{\mathrm{p}}^{+}(m_0)$. The index set $\mathcal{J}$ is given by

$$\mathcal{J} = \left\{m \middle| m_0 \leq m < m_0 + \Delta_{\mathrm{p}}^{+}(m_0)\right\}. \tag{44}$$

Iteration steps (10), (18) and (12) are repeated.

In the following paragraphs we will present some pitch-modification results. In the case of an increased pitch, the pitch periods were first shortened using the method described earlier in this section, after which a number of pitch periods was inserted using the duration-modification method of Section 5. In the case of a decreased pitch, the number of pitch periods was decreased using the duration-modification method of Section 5, after which the pitch periods were extended. The parameters of the duration modification method were chosen to be the same as those in Section 5. The parameters for the pitch-modification method were as follows. The window was a raised cosine, given by (16), with $N_{\mathrm{w}} = 32$. The number of frequency points was given by $N = 128$. The number of iterations was 30.

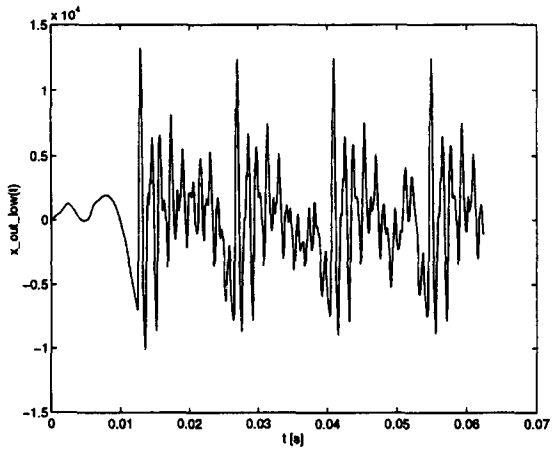


Fig. 22. Vowel /a/ with the pitch reduced by half an octave. Window length $N_w = 32$, number of frequency points $N = 128$, window shifts $S = 1$, number of iterations 30.

Fig. 22 shows 1000 samples of the artificial vowel /a/ of Fig. 2 with the pitch reduced by half an octave, which corresponds to a fraction of 0.71. A low-pitched artificial vowel /a/, generated by feeding an adapted glottal pulse sequence through the vocal tract filter that was used to produce the artificial vowel /a/ of Fig. 2, is shown in Fig. 23. Although the waveforms differ somewhat, there are only very minor audible differences between the two signals. The spectral envelope, characterizing the perceived vowel, is not affected by the pitch modification. This is illustrated in Fig. 24, showing spectral estimates for the original vowel /a/ and its pitch-reduced version, respectively. Figs. 25 and 26 show versions of the Dutch word 'toch', /t ⊃ χ/, with pitches that have been reduced by half an octave and increased by half an octave, respectively.

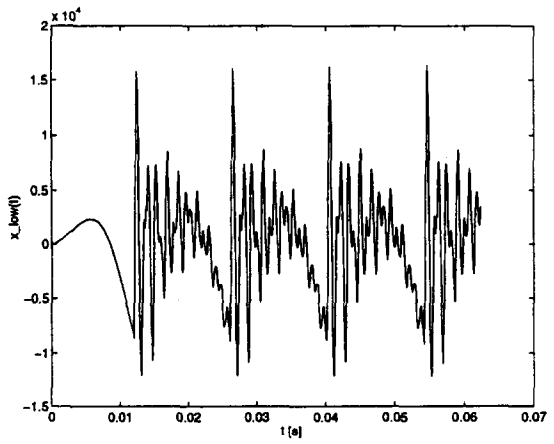


Fig. 23. Low-pitched vowel /a/.

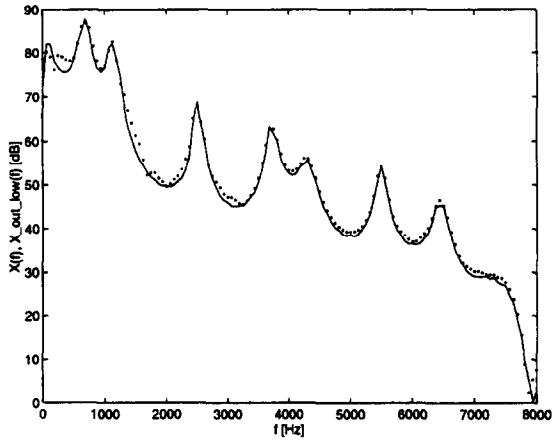


Fig. 24. Solid line: spectrum of vowel /a/. Dotted line: spectrum of vowel /a/ with reduced pitch.

As was the case in Section 5, the quality was judged by informal listening and a comparison was made with the TD-PSOLA method described in Section 5. The short-term signals in which the TD-PSOLA method decomposes the signal were repeated at such a rate that the resulting signal had the desired pitch. A portion was repeated or deleted in order to maintain the correct duration. The listeners

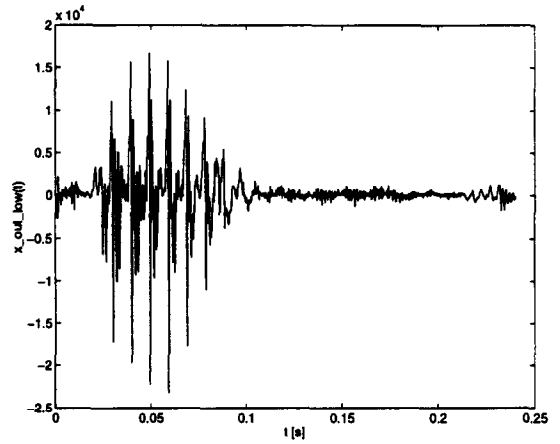


Fig. 25. Dutch word 'toch', /t ⊃ χ/, with the pitch reduced by half an octave. Window length $N_w = 32$, number of frequency points $N = 128$, window shifts $S = 1$, number of iterations 30.

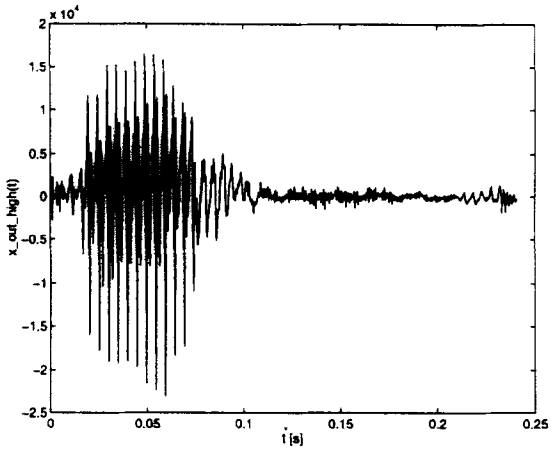


Fig. 26. Dutch word 'toch', /t ɔ χ/, with the pitch increased by half an octave. Window length $N_w = 32$, number of frequency points $N = 128$, window shifts $S = 1$, number of iterations 30.

and the environment were as in Section 5. In the tests, the pitch of the Dutch and English sentences that were also used in Section 5 was changed between a reduction to 30% and an extension to 300% of the original length, for various male as well as female voices [3]. Pitch modifications of between a decrease by an octave and an increase by half an octave were considered to yield good results. Outside this range deteriorations became audible. In addition, nonsense words such as 'fafafe' that are used to extract diphones from were monotonized. The quality was generally better than that obtained with the TD-PSOLA-based method. The quality for female voices seems to improve somewhat if we choose $N_w = 16$, rather than $N_w = 32$.

Informal tests with interfering noise and interfering speech, both with an interference level of -14 dB, lead to the same conclusions as in Section 5, namely that the method seems robust for this type of interference at this level and that it outperforms the TD-PSOLA method.

[3] Examples of sentences used in this test are accessible from the Elsevier WWW server (http://www.elsevier.nl/locate/specome). Signal A is the original sentence "The goose laid an odd egg" uttered by a male speaker at a conversational rate; signal D results from a pitch modification to 150%; signal E results from a pitch modification to 66%.

7. Conclusions

An iterative resynthesis method based on short-time Fourier magnitude only, originally presented in (Griffin and Lim, 1984; Griffin et al., 1984), has been evaluated experimentally. Simulation results indicate how convergence of this method depends on the initial phase, the number of frequency points and the window shift. As a main result it is found that, almost irrespective of the number of frequency points and the window shift chosen, the convergence rate of this method is slow. If we consider the quality of the approximation of short-time Fourier magnitude, the signal-to-noise ratio obtained is around 20 dB. If we consider the quality of the approximation of the original signal, the signal-to-noise ratio is less than 0 dB. The resulting speech is intelligible but noisy. A duration- or pitch-modification method based on modification of the short-time Fourier magnitude is therefore not expected to yield high-quality speech.

Secondly, we have evaluated the same method, but modified in such a way that for every other pitch period the original phase was specified. This leads to a substantial improvement of the quality of the resynthesized signal, especially if the window length is large enough. It is not yet well understood why this improvement occurs or how it depends on the parameters and the pitch period. That will require some more investigations and theoretical analyses that are outside the scope of this paper. The convergence rate is still slow, which means that the method is computationally costly. This changes if the unspecified phase is replaced by the correct phase with an additive error that is uniformly distributed on an interval $[-\alpha\pi, \alpha\pi]$, with $\alpha = 0.2$. In that case the window length N_w, the number of frequency points N and the number of iterations can be reduced. All these steps lead to lower computational costs.

Removing and inserting pitch periods and extending and reducing the lengths of pitch periods can be seen as basic steps in duration and pitch modification of speech signals. The results of the resynthesis of speech signals from their short-time Fourier transform magnitudes with partially specified phases, summarized above, indicate that these basic steps can be done successfully in the short-time Fourier domain.

Duration modification is the simpler of the two

types of modifications as it only involves removing and inserting entire pitch periods. If the signal is unvoiced it is assumed that an (arbitrary) estimate of a pitch period is still available. The point of an insertion or deletion is the position in a pitch period in which spectral change in the time direction according to a certain measure is small. The magnitude of an inserted part can be estimated on the basis of the magnitude in its neighbourhood. An initial phase is computed at the point of the deletion or insertion after which the aforementioned algorithm resynthesizes the speech signal. In the case of voiced speech, the magnitude and phase of the inserted part are chosen in such a way that the periodicity of the signal is maintained. In the unvoiced case the magnitude of the initial part is chosen in such a way that the transition in the time direction is smooth and the initial phase is chosen randomly.

The quality of the duration modification method has been judged by informal listening. Between a reduction to 50% and an extension to 200%, the quality was found to be good. Outside this range some deteriorations became audible. In all cases the method was preferable to a TD-PSOLA method, especially when the time-scale was modified by more than 50% in either direction. There is some dependence on parameter settings. It seems that in the case of female voices better results are obtained with shorter window lengths. We have also done some evaluations with interfering noise and interfering speech. The parameter settings were identical to those used in the case of clean speech. The pitch estimation of the corrupted signal was estimated. In both cases the interference level was -14 dB. The method seems very robust for these types of interferences, and, especially in the case of interfering speech, clearly outperforms the TD-PSOLA method.

Modification of the pitch is more complicated than modification of the duration. When the pitch is decreased by a fraction, the number of pitch periods is decreased by this fraction first and then the length of each pitch period is increased by the same fraction. When the pitch is increased by a certain fraction, the length of each pitch period is decreased by this fraction first and then the number of pitch periods is increased by the same fraction. A pitch period is modified at its end, just before the next pitch period becomes noticeable in the short-time Fourier transform. Measures for locating these positions have been developed. Decreasing the length of a pitch period is similar to removing an entire pitch period, but a shorter part is removed. The phase of the original signal is copied and is allowed to change during the iteration process. Increasing the length of a pitch period is more complicated. Complex linear prediction is used to estimate the inserted part. Only the phase at the point of the insertion is allowed to change during the iteration process.

The quality of the result of the pitch modification-algorithm was also judged by informal listening. Pitch modifications between a decrease by an octave and an increase by half an octave were considered to yield good results. Outside this range deteriorations became audible. The quality was generally found to be better than that obtained with the TD-PSOLA method. The same type of dependence on window length that was observed with the duration modification method was also observed here: shorter windows seem to yield somewhat better results for female voices.

Informal tests with interfering white noise and interfering speech, both with an interference level of -14 dB, led to the same conclusions as those obtained with the duration-modification method, namely that the method seems robust for this type of interference at this level and that it outperforms the TD-PSOLA method.

The observation that in the case of female voices shorter windows yield somewhat better results in modifying duration or pitch led to the idea that there could be a relation between the pitch period and the optimum window length.

Although both duration and pitch modification methods are capable of producing high-quality speech [4], there is one drawback, namely the computational complexity. The methods involve many more

[4] Examples of speech signals modified with the technique here presented are accessible from the Elsevier WWW server (http://www.elsevier.nl/locate/specome). Signal A is the original sentence "The goose laid an odd egg" uttered by a male speaker at a conversational rate; signals B and C were obtained after time-scale modification only; singals D and E were obtained after pitch modification only; signal F was obtained after a combination of pitch modification to 150% and time-scale extension to 150%.

operations than TD-PSOLA. This means that their application areas must lie where quality is of paramount importance and complexity or computation time is not.

Acknowledgements

The authors would like to thank Sylvie Mozziconacci and Ralf Fassel for their translations of the abstract and Paul Moers for his simulation work.

References

B.S. Atal and S.L. Hanauer (1971), "Speech analysis and synthesis by linear prediction of the speech wave", *J. Acoust. Soc. Amer.*, Vol. 50, pp. 637–655.

G. Bailly and C. Benoit, Editors (1992), *Talking Machines, Theories, Models, and Designs* (North-Holland, Amsterdam).

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

D.W Griffin, D.S. Deadrick and J.S. Lim (1984), "Speech synthesis from short-time fourier transform magnitude and its application to speech processing", *Proc. Internat. Conf. Acoust. Speech Signal Process.-84, San Diego.*

C. Hamon, E. Moulines and F. Charpentier (1989), "A diphone synthesis system based on time-domain prosodic modifications of speech", *Proc. Internat. Conf. Acoust. Speech Signal Process.-89, Glasgow*, pp. 238–241.

D.J. Hermes (1988), "Measurement of pitch by subharmonic summation", *J. Acoust. Soc. Amer.*, Vol. 83, No. 1, pp. 257–264.

A.J.E.M. Janssen, R.N.J. Veldhuis and L.B. Vries (1986), "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 34, No. 2, pp. 317–330.

S.L. Marple, Jr. (1987), *Digital Spectral Analysis with Applications* (Prentice Hall, Englewood Cliffs, NJ).

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, December, pp. 453–467.

D. O'Shaughnessy (1990), *Speech Communication* (Addison-Wesley, Reading, MA).

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

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. 30, No. 4, pp. 566–578.

W. Verhelst and M. Roelands (1993), "An overlap-add technique based on waveform similarity (WSOLA) for high-quality time-scale modification of speech", *Proc. Internat. Conf. Acoust. Speech Signal Process.-93, Minneapolis*, pp. II-554–II-558.