ELSEVIER

Speech Communication 30 (2000) 207–221

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Overlap-add methods for time-scaling of speech ☆

Werner Verhelst [1]

Ku Leuven – ESAT, Kardinaal Mercierlaan 94, B-3001 Heverlee, Belgium

Received 16 July 1998; received in revised form 17 May 1999; accepted 14 September 1999

Abstract

In this tutorial on time-scaling we follow one particular line of thought towards computationally efficient high quality methods. We favor time-scaling based on time–frequency representations over model based approaches, and proceed to review an iterative phase reconstruction method for time-scaled magnitude spectrograms. The search for a good initial phase estimate leads us to consider synchronized overlap-add methods which are further optimized to eventually arrive at WSOLA, a technique based on a waveform similarity criterion. © 2000 Elsevier Science B.V. All rights reserved.

Zusammenfassung

In diesem Tutorium über Zeitverlaufsskalierung wird ganz besonders der Gedanke an ein aufwandsgünstiges und hochqualitatives Verfahren verfolgt. Wir favorisieren eine Zeitverlaufsskalierung auf der Basis von Zeit–Frequenz-Darstellungen vor modellbasierten Ansätzen und betrachten eine iterative Phasenrekonstruktionsmethode für zeitskalierte Spektrogramme des Amplitudenbetrags. Die Suche nach einer geeigneten initialen Phase führt uns zu einer Betrachtung von synchronisierten Overlap-Add Verfahren, die bei weiterer Optimierung schließlich zu WSOLA führen, eine Technik, die auf dem Kriterum der Signalfosrmähnlichkeit beruht.□ © 2000 Elsevier Science B.V. All rights reserved.

Résumé

Dans cet exposé sur la modification de la structure temporelle du signal de parole, nous opterons pour l'utilisation des représentations temps–fréquence du signal, plutôt que pour des représentations par modèles. Nous examinerons une méthode itérative permettant de reconstruire une fonction de phase pour spectrogrammes d'amplitude modifiés. La recherche d'une bonne condition initiale pour démarrer l'itération nous ammènera aux méthodes de recouvrement-addition synchronisées et notamment à WSOLA, une technique basée sur un critère de ressemblance entre formes d'ondes. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Speech processing; Speech modification; Time-scaling; Time-warping; Short-time Fourier transform; Overlap-add; WSOLA

1. Introduction

Methods that allow the duration of speech to be modified can be used to create useful functions for many products. Shortening the duration of original speech messages allows for corresponding

☆ Speech files available. See www.elsevier.nl/locate/specom.

[1] Parts of this work were performed at the Institute for Perception Research, Eindhoven and at the Vrije Universiteit Brussel.

E-mail address: werner.verhelst@esat.kuleuven.ac.be (W. Verhelst).

0167-6393/00/$ - see front matter
PII: S0167-6393(99)00051-5

savings in storage and transmission. Rendering speech at a rate that can be chosen arbitrarily different from the original rate can increase the ease-of-use and the efficiency of speech reproduction equipment. It can allow, for example, for faster listening to messages recorded on answering machines, voice mail systems, information services, etc. or for synchronizing speech from dictation with the typing speed.

The problem with time-scaling a speech signal $x_a(t)$ of original duration Δt lies with the corresponding frequency distortion. It is common experience that when $x_a(t)$ is played back at a higher speed than the recording speed, the resulting sound is distorted in that its pitch is raised. Conversely, when the recording is played back at a lower speed the pitch is lowered. Not only pitch is affected but timbre is distorted as well, such that when the playback speed differs significantly from the recording speed comprehension of the messages becomes difficult or even impossible.

The duality between time-scaling and frequency-scaling becomes mathematically clear by considering the signal $y_a(t)$ that corresponds to an original signal $x_a(t)$ played at a speed α times higher than the recording speed. Thus, an original time span Δt is played in $\Delta t/\alpha$ and $y_a(t) = x_a(\alpha t)$. From the definition of the Fourier transformation for analog signals, we find that uniform scaling in one domain corresponds to the reverse scaling in the transformed domain: [2]

$$y_a(t) = x_a(\alpha t) \leftrightarrow Y_a(\Omega) = \frac{1}{|\alpha|} X_a\left(\frac{\Omega}{\alpha}\right).$$

As discussed in more detail in Section 2, the intended time-scaling clearly does not correspond to the mathematical time-scaling. We rather require a scaling of the *perceived* timing attributes, such as speaking rate, without affecting the perceived frequency attributes, such as pitch. Because of the mathematical duality between time domain and frequency domain representations, we can consider two equivalent formulations for this problem:

(I) Modify the time domain representation of signal $x_a(t)$ without altering its perceived frequency attributes.

(II) Modify the frequency domain representation of signal $y_a(t)$ without altering its perceived time structure.

In this paper we will consider the problem in its first formulation, and discuss how we can construct a time-scaled signal that does not suffer perceived frequency distortion. At the end of the paper we will present an example application that uses time-scaling methods to solve the problem stated in its dual form (II). There we will consider time-varying playback of analog recordings, where an amount of information (phonemes per second) is played at a given target speed, such that the available analog signal $y_a(t)$ already has the desired duration but needs to be frequency-scaled to restore the original pitch and timbre.

2. General considerations on time-scaling

2.1. The time-scaling function

Formally, time-scale modifications are specified by defining a mapping $n \rightarrow n' = \tau(n)$ between the original time-scale and the modified time-scale. This mapping defines the so-called *time-scaling* or *time-warping* function (Moulines and Verhelst, 1995). It specifies that the sounds which occur at time n in the original signal should occur at time n' in the time-scaled signal. Sometimes, a time-varying time-modification rate $\beta(t) > 0$ is specified, from which the time-scaling function can be derived as

$$n \rightarrow n' = \tau(n) = \frac{1}{T} \int_0^{nT} \beta(u)\,\mathrm{d}u, \tag{1}$$

where T is the sampling period. At time instances where $\beta(t) > 1$, the time-scaling corresponds to slowing-down the original; at time instances where $0 < \beta(t) < 1$, the time-scaling corresponds to speeding-up the original.

The appropriate shape for the time-warping function depends on the application. In diphone synthesis, for instance, it follows from the inherent

[2] The factor $1/|\alpha|$ is a consequence of energy relations between the original and the scaled signal.

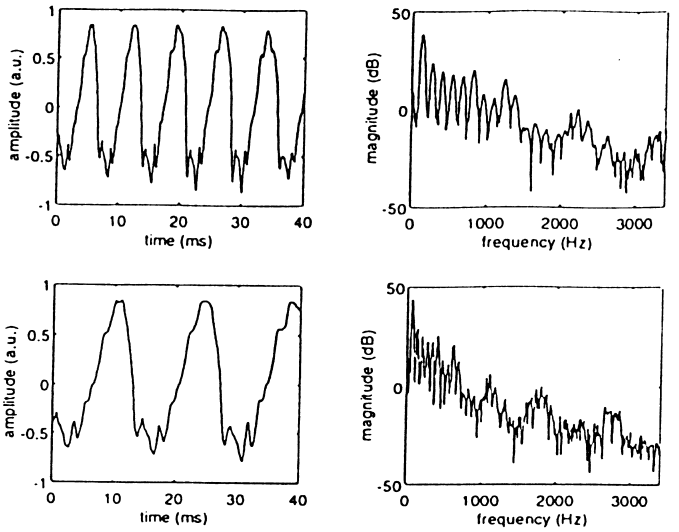

Fig. 1. Illustration of the duality between time domain and frequency domain. The upper row shows a 40 ms voiced speech segment and its spectrum; the second row illustrates that when this signal is played at half speed it is stretched twofold in the time domain and compressed twofold in the frequency domain.

duration of the elements in the speech database and their desired duration in the output signal as computed from a timing model in the intonation module. In musical instrument synthesis, depending on the type of instrument, the corresponding sound can be played with a constant duration or can be sustained until the performer releases the note's key. In both examples, the time-scaling function is nonlinear as different portions of the speech or music are scaled to different extend. Linear time-scaling functions are appropriate for overall tempo scaling.

2.2. Time and frequency patterns of sound

In Section 1, we showed that when a signal is scaled along the time axis by a certain factor, its Fourier representation is scaled along the frequency axis by the inverse of this factor (see Fig. 1). Such a signal would be perceived as a time-scaled and frequency-modified version of the original, and this does not correspond to our intuitive expectation related to hearing a time-scaled version of an original acoustic signal. This is because it is our most elementary experience that sound has a time pattern as well as a frequency pattern (Gabor, 1947) and that these patterns are relatively independent as they are related to the rhythm and the melody, respectively. We therefore required a different type of scaling in Section 1 (one that does not affect the perceived frequency attributes of the signal).

The perceived time and frequency attributes of acoustical signals are related to their physical properties in rather complicated ways, see for example (Moore, 1982). [3] Consider for example a simple acoustical signal, consisting of a finite portion of a pure tone (sine wave). In general, this signal could be perceived to consist of up to three consecutive acoustic events:

- The signal onset could be perceived as a transient event that is called attack (if the segment starts in a relatively abrupt way) or as a tone that gradually emerges from background silence (if a sufficiently slow fade-in is applied). Thus, in this case, the physical time domain properties of the signal control whether or not a frequency

[3] Psychoacoustics is the scientific discipline that studies the relationships between acoustical properties and perceived aspects of sounds. It is part of the larger field of psychophysics, see for example (Lindsay and Norman, 1977).

domain attribute of the sound (tone-pitch) can be perceived. Actually, time domain modifications to the attack envelope could alter the perceived timbre of the tone as well (Roads, 1998).

- The middle portion of the signal would be perceived as a sustained pure tone of finite duration with a pitch that is a function of the sine wave's frequency. Thus, in this case, it is a physical frequency domain property which controls the perceived attributes of the sound. Increasing the sine wave's frequency increases the perceived pitch, but changes the loudness as well: from about 2.5 kHz onwards the loudness will gradually decrease until the tone vanishes completely at about 20 kHz. Thus, the variation of the physical frequency domain parameter controls a perceived time domain parameter as well, i.e., the loudness contour.
- The signal offset could be perceived as a transient sound or as a more gradual tone release, similar to the signal onset case.

This simple example illustrates that the physical time pattern of acoustical signals can affect the perceived frequency structures of sound, and that the physical frequency structure can affect the perceived time pattern of sound. The goal of time-scaling is to modify the perceived time pattern of speech in accordance with the specified time-scaling function. We could say that we want the time-scaled version of an acoustic signal to be perceived as the same sequence of acoustic events from the original signal being reproduced according to a scaled time pattern.

2.3. Approaches to time-scaling of speech

In the case of speech signals, we have a number of synthesis models that can be used to produce speech from a set of production parameters. One general approach for time-scaling speech could therefore consist of first analyzing the original speech signal in order to obtain the time-varying vector of production parameters $\vec{p}_x(n)$, then applying the desired time-scale transformation to the production parameters $\vec{p}_y(n) = \vec{p}_x(\tau^{-1}(n))$, and synthesizing the corresponding time-scaled signal. Obviously this strategy will work provided that the vector of model parameters $\vec{p}_x(n)$ is an adequate description for all perceptually important frequency domain aspects of the signal at each particular time instant $n = n_0$. An LPC vocoder is one example of a system that can be used to time-scale speech signals according to this strategy. Instantaneous speech coders on the other hand are not suited since by definition their code word $c(n)$ at a particular time instant $n = n_0$ does not carry sufficient information to decide about voicing, pitch or timbre (a number of consecutive code words are needed together to derive that information).

In selecting an appropriate analysis–synthesis model for time-scaling a trade-off would have to be made between computational complexity and speech quality and it is likely that it will be hard to strike a good compromise with this parametric type of approach. Model based approaches can be interesting for coding where the value of speech production parameters $\vec{p}_x(n)$ can be efficiently quantized, or for synthesis where they are modified according to certain synthesis rules. In time-scaling, however, we have no real use for knowledge about the value of production parameters as we essentially want to reproduce exactly the same sounds in our time-scaled result. In that case the selection of a production model that is sufficiently accurate to describe speech with high acoustic fidelity and that would at the same time remain computationally attractive for most practical applications could be hard.

In this paper, we will concentrate on non-parametric approaches for time-scaling and use the criterion that a time-scaled signal should be perceived to consist of the same sequence of acoustic events as the original signal but with a modified timing structure. Since sounds are perceived to have frequency domain features like pitch and timbre that evolve over time, non-parametric approaches can use a time–frequency representation $X(\omega, n)$ for the speech signal $x(n)$, in which the perceived attributes of $x(n)$ at a given instant $n = n_0$ are to be represented in $X(\omega, n_0)$ along the frequency dimension. Time-scaling would then be achieved by applying the desired transformation to the time axis $\widehat{Y}(\omega, n) = X(\omega, \tau^{-1}(n))$ and transforming the result back to pure time domain.

Such methods differ in their choice of time–frequency representation (i.e., in their analysis strategy), but because current time–frequency representations only achieve an approximate separation of the timing structure from other perceived aspects of sound, they also differ in their choice of modification strategy $\widehat{Y}(\omega,n) = M_{xy}[X(\omega,n)]$. Moreover, as $\widehat{Y}(\omega,n)$ is constructed as a two-dimensional representation for a one-dimensional signal, it often occurs that $\widehat{Y}(\omega,n)$ is not a valid time–frequency representation in that no signal $\widehat{y}(n)$ exists that has $\widehat{Y}(\omega,n)$ as its time–frequency transform. Therefore time-scaling methods additionally differ in the synthesis strategy that is adopted (i.e., in how a signal $y(n)$ with transform $Y(\omega,n)$ is constructed from the desired, but possibly invalid, transform $\widehat{Y}(w,n)$). Roucos and Wilgus (1985) and Verhelst and Roelands (1993) describe examples of time-scale modification algorithms that are based on linear time–frequency representations; Griffin and Lim (1984) for example uses a quadratic time–frequency representation, and d'Alessandro (1989) uses a specially crafted representation that is inspired by acoustic perception theory. In this paper we have chosen to restrict ourselves to time-scaling methods that use the short-time Fourier transformation and overlap-add construction methods because in our opinion they currently achieve a very attractive quality at a very low computational cost. Before considering the application of time-scaling, we will review the basic techniques of short-time Fourier transformation and OLA synthesis in the next section, and show how OLA synthesis can be used to compute the inverse short-time Fourier transform.

3. The short-time Fourier transform and overlap-add synthesis

3.1. The short-time Fourier transform

The Fourier transform

$$X(\mathrm{e}^{\mathrm{j}\omega}) = \sum_{n=-\infty}^{+\infty} x(n)\,\mathrm{e}^{-\mathrm{j}\omega n}$$

is the most important frequency domain representation for stationary signals (whose characteristics do not change with time). If we consider speech as a signal with slowly evolving characteristics (i.e., as a quasi-stationary signal), we can apply a short-time analysis strategy together with Fourier transformation to obtain the so-called short-time Fourier transform (STFT) as the desired time–frequency representation (Deller Jr. et al., 1993). We define the STFT of a signal $x(n)$ by segmenting the signal

$$x_{\mathrm{w}}(n,m) = w(n)x(n+m) \tag{2}$$

and taking the Fourier transform

$$X(\omega,m) = \sum_{n=-\infty}^{+\infty} x(n+m)w(n)\,\mathrm{e}^{-\mathrm{j}\omega n}. \tag{3}$$

A conceptual disadvantage of this approach towards time-varying analysis is that the analysis precision is limited by the windowing operation and non-stationarity; a practical advantage is that short-time analysis works with consecutive, possibly overlapping, signal segments and is easily amenable to on-line processing.

3.2. The overlap-add synthesis method

As noted earlier, by modifying $X(\omega,n)$ (to achieve time-scaling in this case) the result may no longer represent an STFT in that a signal which has the modified transform $\widehat{Y}(\omega,n)$ as its STFT may not exist. Still $\widehat{Y}(\omega,n)$ would contain the information which best characterizes the signal modification we had in mind, such that a special synthesis formula is required which leads to the correct result if $\widehat{Y}(\omega,n)$ is an STFT and to a reasonable result otherwise. One such synthesis method uses overlap-addition (OLA). As introduced by Griffin and Lim (1984) this method constructs $y(n)$ such that its STFT $Y(\omega,n)$ is maximally close to $\widehat{Y}(\omega,n)$ in least squares sense, i.e., such that the total squared error

$$E = \sum_k \frac{1}{2\pi} \int_{-\pi}^{+\pi} \left|\widehat{Y}(\omega,k) - Y(\omega,k)\right|^2 \mathrm{d}\omega \tag{4}$$

is minimized over all signals $y(n)$ (the sum is over all time instants k for which $\widehat{Y}(\omega,k)$ is defined). From Parseval's theorem, Eq. (4) can be written as

$$E=\sum_k\sum_{m=-\infty}^{+\infty}\left(\widehat{y}_{\mathrm{w}}(m,k)-y(m+k)w(m)\right)^2,$$

where $\widehat{y}_{\mathrm{w}}(m,k)$ is the inverse Fourier transform of $\widehat{Y}(\omega,k)$. The signal $y(n)$ which minimizes E is obtained by solving

$$\frac{\partial E}{\partial y(n)}=-2\sum_k\left(\widehat{y}_{\mathrm{w}}(n-k,k)-y(n)w(n-k)\right)\times w(n-k)=0.$$

Thus,

$$y(n)=\frac{\sum_k w(n-k)\widehat{y}_{\mathrm{w}}(n-k,k)}{\sum_k w^2(n-k)},\tag{5}$$

where

$$\widehat{y}_{\mathrm{w}}(n-k,k)=\frac{1}{2\pi}\int_{-\pi}^{+\pi}\widehat{Y}(\omega,k)\,\mathrm{e}^{\mathrm{j}\omega(n-k)}\,\mathrm{d}\omega$$

is the inverse Fourier transform of $\widehat{Y}(\omega,k)$ delayed by k samples.

To see how this OLA synthesis operates, consider the operations of STFT and inverse STFT of a signal $x(n)$ using Eqs. (3) and (5) with an analysis window $w(n)$ centered around time origin.

- To compute $X(\omega,m)$, the signal is advanced m points in time and windowed to obtain $x_{\mathrm{w}}(n,m)=x(n+m)w(n)$ (Eq. (2)). The STFT $X(\omega,m)$ is then obtained by taking the Fourier transformation of $x_{\mathrm{w}}(n,m)$ towards n (Eq. (3)).
- To obtain the inverse STFT, the inverse Fourier transform of $X(\omega,m)$ is computed to recover the windowed segments $x_{\mathrm{w}}(n,m)$. This result is windowed again using the synthesis window [4] to obtain $w(n)x_{\mathrm{w}}(n,m)$. As all these segments were positioned around time origin during the analysis, they now have to be delayed to move each one back to its original location along the time axis (i.e., around time k for segment number k). The result is then obtained by summing all these segments and dividing by a time-varying normalization weight (Eq. (5)):

$$y(n)=\frac{\sum_k w(n-k)x_{\mathrm{w}}(n-k,k)}{\sum_k w^2(n-k)}=\frac{\sum_k w^2(n-k)x(n)}{\sum_k w^2(n-k)}=x(n).$$

Thus, the OLA [5] synthesis formula reconstructs the original signal if $X(\omega,m)$ is a valid STFT, [6] or constructs a signal whose STFT is maximally close to $X(\omega,m)$ in least squares sense otherwise. Additionally it can be noted that the denominator in Eq. (5) is actually needed only to compensate for a possible non-uniform weighting of samples in the windowing procedure. Also, the synthesis operation can be simplified if the windowing function and the synthesis time instants k can be chosen such that

$$\sum_k w^2(n-k)=1.\tag{6}$$

A common choice in speech processing that satisfies this simplifying condition is a hanning window with 50% overlap between successive segments; some other possibilities are listed in (Griffin and Lim, 1984).

4. OLA time-scaling

It can be noted that with OLA synthesis we are close to realizing time-scale modifications using time domain operations only. In fact we can see that, by adopting a short-time analysis strategy for constructing $X(\omega,m)$ and by using the OLA criterion for synthesizing a signal $y(n)$ from the modified representation $\widehat{Y}(\omega,m)=M_{xy}[X(\omega,m)]$, we will always obtain modification algorithms that can be operated in the time domain if the modifi-

[4] The synthesis window need not necessarily be the same as the analysis window, but this choice usually makes good sense in practice.

[5] The formula was initially called LSEE MSTFT, which stands for least squares error estimation from modified STFT (Griffin and Lim, 1984) but is now usually referred to as (a variant of) the OLA method.

[6] Provided that each sample $x(n)$ lies in at least one segment or, equivalently, $\sum_k w^2(n-k)\neq 0\ \forall n$ which we implicitly assumed in our derivation.

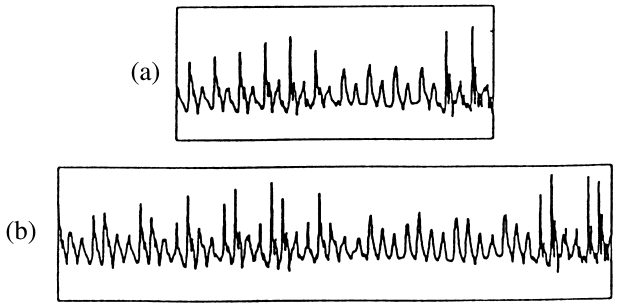

Fig. 2. OLA synthesis from the time-scaled STFT does not succeed to replicate the quasi-periodic structure of the signal (a) in its output (b).

cation operator $M_{xy}[\cdot]$ works on the time index m only:

$$\widehat{Y}(\omega,m) = X(\omega, M_{xy}[m]) \quad \text{(modification)},$$
$$y_{\mathrm{w}}(n,m) = x_{\mathrm{w}}(n, M_{xy}[m]) \quad \text{(inverse FT)},$$
$$y(n) = \frac{\sum_m w(n-m)x_{\mathrm{w}}(n-m, M_{xy}[m])}{\sum_m w^2(n-m)}$$
(OLA synthesis).

In that case we see from the last equation above that the modification is obtained by excising segments $x_{\mathrm{w}}(n, M_{xy}[m])$ from the input signal and repositioning them along the time axis before constructing the output signal by weighted overlap-addition of the segments. However, as illustrated in Fig. 2, if we hurry to apply the above formula for realizing a time-warp $\tau(m)$, poor results will generally be obtained when using $\widehat{Y}(\omega,m) = X(\omega,\tau^{-1}(m))$.

It is the same short-time analysis principle which lead us to hope for time-domain implementations that now causes the problems by constructing a two-dimensional representation $x_{\mathrm{w}}(n,m) = w(n)x(n+m)$ in which the two time-scales are not independent such that important information about the time structure of the signal $x(n)$ is represented both in ω and in m. Consider, for example, the case of a periodical signal $x(n) = x(n+N)$. If the window is sufficiently long, each segment $x_{\mathrm{w}}(n,m)$ contains several periods of the original. At the same time, this periodic structure also exists among different segments $x_{\mathrm{w}}(n,m+N) = x_{\mathrm{w}}(n,m)$ (all segments separated by a multiple of the period N are identical). By arbitrarily repositioning these segments, as in $\widehat{Y}(\omega,m) = X(\omega,\tau^{-1}(m))$, we destroy the relationship between the time structure inside the segments and the time structure across the segments. In the example shown in Fig. 2 this leads to the quasi-periodic structure of the input speech signal (a) not being preserved in the time-scaled output (b). (In this example the attempted time-scaling consisted of a reduction of the apparent speaking rate to 60% of the original.)

The phase component $\Phi(\omega,m)$ of the complex STFT $X(\omega,m)$ carries information about the signal's time structure inside the analysis window. [7] A better separation with information on perceptual characteristics in ω and time structural information in m is found in the spectrogram $|X(\omega,m)|$ where each magnitude spectrum shows pitch information in its harmonic structure and formant information in its spectral envelope. Because it contains no phase information, a magnitude spectrum does not specify the precise time structure of the signal segment $x_{\mathrm{w}}(n,m)$ nor does it carry information concerning the position of the signal $x(n)$ relative to the window $w(n)$. If a sufficiently long window is used (several pitch periods long), it was shown in (Griffin and Lim, 1984) that good quality time-scaled signals can be obtained from the time-scaled spectrogram

[7] The group delay $\tau_g(\omega,m) = -\mathrm{d}\Phi(\omega,m)/\mathrm{d}\omega$ determines the relative position around time m of the signal's energy at frequency ω.

$$|\widehat{Y}(\omega, kS)| = |X(\omega, \tau^{-1}(kS))|,$$

where S is a downsampling factor introduced to reduce the amount of information that needs to be processed.

In order to construct $y(n)$, however, the phase of the STFT $\widehat{Y}(\omega, kS)$ needs to be constructed from its magnitude $|\widehat{Y}(\omega, kS)|$. This can be done as shown by Griffin and Lim, as follows.

- An initial phase function $\Phi_0(\omega, kS)$ is guessed (e.g., by random choice).
- At each iteration step, $i = 1, \ldots$
  1. OLA is used to form the ith signal estimate $y_i(n)$ from the specified spectrogram $|\widehat{Y}(\omega, kS)|$ and the current phase guess

$$y_i(n) = \frac{\sum_{k=-\infty}^{+\infty} w(n - kS) \frac{1}{2\pi} \int_{-\pi}^{+\pi} |\widehat{Y}(\omega, kS)| \, e^{j\Phi_{i-1}(\omega, kS)} e^{j\omega(n-kS)} \, d\omega}{\sum_{k=-\infty}^{+\infty} w^2(n - kS)}.$$

  2. The ith phase function $\Phi_i(\omega, kS)$ is obtained from the STFT $Y_i(\omega, kS)$ of the current signal estimate $y_i(n)$.

Griffin and Lim showed that this algorithm converges in that the distortion measure

$$D_M(i, \Phi_0(\omega, kS)) \triangleq \sum_{k=-\infty}^{+\infty} \frac{1}{2\pi} \int_{-\pi}^{+\pi} \left(|\widehat{Y}(\omega, kS)| - |Y_i(\omega, kS)| \right)^2 d\omega$$

decreases with each iteration step. However, the method does not necessarily converge to the global optimum, and convergence is rather slow. It can also be noted that each iteration step is rather computationally intensive: this approach to time-scaling needs to modify the STFT along the frequency dimension (by modifying phase) and does not operate in time domain only.

5. Synchronized OLA time-scaling

Since OLA time-scaling is an iterative procedure that slowly converges to a local optimum it becomes important that a good initial estimate $\Phi_0(\omega, kS)$ or, equivalently, a good choice for $y_1(n)$ can be proposed. Roucos and Wilgus (1985) experimentally studied the convergence of OLA time-scaling and found that for initial estimates like Gaussian white noise 100 iterations were typically required to obtain high quality results. In their effort to find a better initial estimate that would significantly reduce the required number of iterations, Roucos and Wilgus (1985) proposed a construction method for $y_1(n)$ whose result has by itself already such high quality that further iterations are no longer needed and do not in fact improve the subjective speech quality. This algorithm for time-scaling is called the synchronized overlap-add method (SOLA) and can be described in the following.

As discussed in Section 4, straightforward OLA synthesis from the time-scaled and downsampled STFT $\widehat{Y}(\omega, kS) = X(\omega, \tau^{-1}(kS))$ results in a signal

$$y_1(n) = \frac{\sum_k w^2(n - kS) x(n - kS + \tau^{-1}(kS))}{\sum_k w^2(n - kS)}$$

that is heavily distorted, as we illustrated in Fig. 2. Interpreted as a possible initial estimate for iterative OLA time-scaling, $y_1(n)$ corresponds to an initial phase $\Phi_0(\omega, kS)$ that is chosen equal to the actual phase of the individual segments $x_w(n, \tau^{-1}(kS))$. This choice would certainly be fine for the individual segments but, as we discussed in Section 4, repositioning the segments from their original time position $\tau^{-1}(kS)$ to the required synthesis time position kS destroys the original phase relations across segments. Roucos and Wilgus noted that this repositioning of segments corresponds to the introduction of a linear phase difference between the individual segments that disrupts the periodical structure of the voiced parts of speech, unless the difference would be equal to a multiple of the pitch period. With their SOLA algorithm (Roucos and Wilgus, 1985) they propose to avoid pitch period discontinuities at waveform segment boundaries by realigning each input segment to the already formed portion of the output signal before performing the OLA operation. Thus, SOLA constructs the time-scale modified signal

$$y(n) = \frac{\sum_k v(n - kS + \Delta_k) x(n + \tau^{-1}(kS) - kS + \Delta_k)}{\sum_k v(n - kS + \Delta_k)} \tag{7}$$

in a left-to-right fashion with a windowing function $v(n)$, and with shift factors Δ_k that are chosen such as to maximize the cross-correlation coefficient between the current segment $v(n-kS+\Delta_k)x(n+\tau^{-1}(kS)-kS+\Delta_k)$ and the already formed portion of the output signal

$$y(n;k-1) = \frac{\sum_{l=-\infty}^{k-1} v(n-lS+\Delta_l)x(n+\tau^{-1}(lS)-lS+\Delta_l)}{\sum_{l=-\infty}^{k-1} v(n-lS+\Delta_l)}.$$

SOLA is computationally efficient since it requires no iterations and can be operated in the time domain. As discussed earlier, time domain operation implies that the corresponding STFT modification affects the time axis only. In case of SOLA, we have

$$\widehat{Y}(\omega, kS-\Delta_k) = X\big(\omega, \tau^{-1}(kS)\big).$$

The shift parameters Δ_k thus imply a tolerance on the time warp function: in order to ensure a synchronzed overlap-addition of segments, the desired time-warp function $\tau(n)$ will not be realized exactly. A deviation on the order of a pitch period should be allowed.

Several alternative synchronization methods can be constructed. The pitch-synchronous overlap-add method (PSOLA (Moulines and Charpentier, 1990)), for example, specifies that the input segments $x_w(n,m)$ are to be chosen pitch-synchronously and are likewise to be overlap-added in a pitch- synchronized way. Thus,

$$\widehat{Y}(\omega, S_k) = X(\omega, \tau^{-1}(S_k)+\Delta_k)$$

with S_k synthesis pitch marks, and timing tolerance Δ_k such that $\tau^{-1}(S_k)+\Delta_k$ is an analysis pitchmark.

6. WSOLA: an overlap-add technique based on waveform similarity

6.1. Efficient synchronized OLA time-scaling

From the above discussions we observe that, in order to construct an efficient high-quality time-scaling algorithm based on OLA, a tolerance Δ_k on the precise time-warp function that will be realized is needed to allow a synchronized overlap-addition of original input segments to be performed in the time domain. This tolerance can be used like in SOLA to realize segment synchronization during synthesis $\widehat{Y}(\omega, kS-\Delta_k) = X(\omega, \tau^{-1}(kS))$. However, as the Δ_k are not known beforehand, the denominator in the OLA formula (7) could not be made constant in that case. A further reduction of computational cost would be possible by using fixed synthesis time instants $S_k = kS$ and a window $v(n)$ such that $\sum_k v(n-kS) = 1$. Proper synchronization must then be ensured during the segmentation

$$\widehat{Y}(\omega, kS) = X(\omega, \tau^{-1}(kS)+\Delta_k). \tag{8}$$

Thus it would seem that a most efficient realization of OLA time-scaling would use the simplified synthesis equation

$$y(n) = \sum_k v(n-kS)x(n+\tau^{-1}(kS)-kS+\Delta_k),$$

where Δ_k are chosen such as to ensure sufficient signal continuity at waveform segment boundaries according to some criterion. WSOLA (Verhelst and Roelands, 1993) proposes a synchronization strategy inspired on a time-scaling criterion.

6.2. A waveform similarity criterion for time-scaling

We considered that a time-scaled version of an original signal should be perceived to consist of the same acoustic events as the original signal but with these events being produced according to a modified timing structure. In WSOLA we assume that this can be achieved by constructing a synthetic waveform $y(n)$ that maintains maximal local similarity to the original waveform $x(m)$ in all neighborhoods of related sample indices $m = \tau^{-1}(n)$. Using the symbol $(=)$ to denote maximal similarity and using the window $w(n)$ to select such neighborhoods, we require

$$\forall m : y(n+m)w(n) \quad (=)x(n+\tau^{-1}(m)+\Delta_m)w(n) \tag{9}$$

or equivalently

$$\forall m : \ \widehat{Y}(\omega, m)(=)X(\omega, \tau^{-1}(m)+\Delta_m). \tag{10}$$

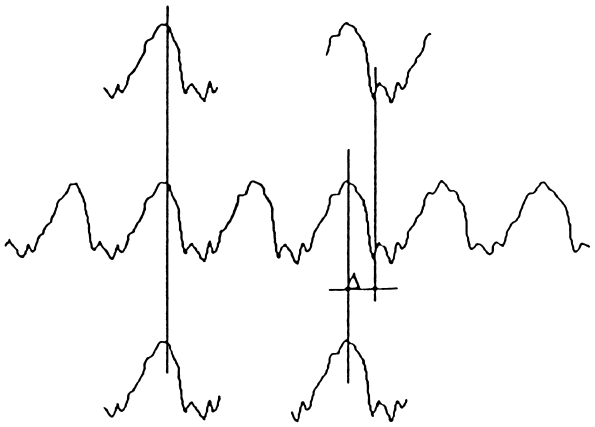

Fig. 3. Alternative interpretation of timing tolerance parameters Δ.

Comparing Eqs. (9) and (10) with Eq. (8), we find an alternative interpretation for the timing tolerance parameters Δ_k as we see that the waveform similarity criterion and the synchronization problem are closely related. As illustrated in Fig. 3, the Δ_m in Eqs. (9) and (10) were introduced because in order to obtain a meaningful formulation of the waveform similarity criterion, two signals need to be considered identical if they only differ by a small time-offset. [8] Referring to Fig. 3, we need to express that the waveshapes of segments from the quasi-periodic signal in the middle of the figure are similar at all time instants. Such similarity goes unnoticed in the upper pair of segments because they are located at different positions in their respective pitch cycles. By introducing a tolerance Δ_m on the time instants around which segment waveforms are to be compared, the quasi-stationarity of the signal can easily be detected from the lower pair of segments that was synchronized by letting $\Delta_m = \Delta$. We can thus conclude that using the requirement of waveform similarity between input and output signals as a criterion for time-scaling, readily implies that a synchronization of input and output segments will have to take place.

[8] It can be noted that waveform similarity was used to approximate sound similarity. Because two signals that differ only by some time-offset sound the same, we also need to declare their waveforms to be similar.

As Eq. (8) can be viewed as a downsampled version of Eq. (10), we propose to select the parameters Δ_k such that the resulting time-scaled signal

$$y(n) = \sum_k v(n - kS)x(n + \tau^{-1}(kS) - kS + \Delta_k) \tag{11}$$

maintains maximal local similarity to the original waveform $x(m)$ in corresponding neighborhoods of related sample indices $m = \tau^{-1}(n)$.

6.3. WSOLA algorithms

Based on this idea a variety of practical implementations can be constructed. A common version of WSOLA uses a 20 ms hanning window with 50% overlap ($S = 10f_s$, with f_s the sampling frequency in kHz) to construct the signal of Eq. (11) in a left-to-right manner as illustrated in Fig. 4.

Assume the segment labeled (1) in Fig. 4 was the previous segment that was excised from the input signal and overlap-added to the output at time instant $S_{k-1} = (k-1)S$, i.e., synthesis segment (a) = input segment (1). At the next synthesis position $S_k = kS$ we need to chose a synthesis segment (b) that is to be excised from the input around a time instant $\tau^{-1}(S_k) + \Delta_k$, where $\Delta_k \in [-\Delta_{\max} \ldots \Delta_{\max}]$ is to be chosen such that the resulting portion of $y(n)$, $n = S_{k-1}, \ldots, S_k$ will be similar to a corresponding portion of the input. As

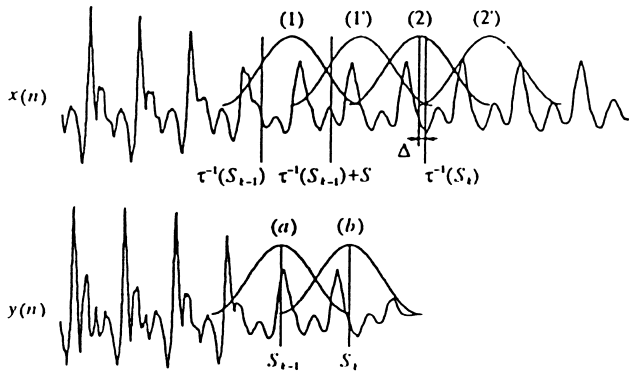

Fig. 4. Illustration of WSOLA time-scaling.

segment $(1')$ overlap-adds with (1) to reconstruct a portion of the original signal $x(n)$ this segment $(1')$ would also overlap-add with segment (a) to reconstruct that same portion of the original in the output signal $y(n)$ (remember that segment (a) = segment (1)). While we cannot accept segment $(1')$ as a legal candidate for synthesis segment (b) if it does not lie in the timing tolerance region $[-\Delta_{\max} \ldots \Delta_{\max}]$ around $\tau^{-1}(S_k)$, we can always use it as a template to select segment (b) such that it resembles segment $(1')$ as closely as possible and is located within the prescribed tolerance interval around $\tau^{-1}(S_k)$ in the input signal. The position of this best segment (2) can be found by maximizing a similarity measure between the sample sequence underlying segment $(1')$ and the input signal. After overlap-addition of synthesis segment (b) = input segment (2) to the output we can proceed to the next synthesis time instant using segment $(2')$ as our next template.

WSOLA proposed the criterion of waveform similarity as a substitute for the time-scaling criterion which required that at all corresponding time instants the original and the time-scaled signal should sound similar. Clearly waveform similarity can only be a valid substitute for sound similarity if the similarity can be made sufficiently close. For time-scaling of speech signals, which are largely made up from long stretches of quasi-periodic waveforms and noiselike waveshapes, a strategy like WSOLA is indeed able to produce close waveform similarity. In that case the precise similarity measure selected does not matter too much in that any reasonable distance measure (like cross-correlation, cross-AMDF, etc.) will do. As illustrated in Figs. 5 and 6, original and WSOLA time-scaled waveforms do indeed show a very close similarity. [9] Also, many variants of the basic technique can be constructed by varying the window function, the similarity measure, the portion of the original $x(n)$ that is to serve as a reference for natural signal continuity across OLA segment boundaries, etc. As many such variants all provide a similar high quality (Verhelst and Roelands, 1993) this design flexibility can be used to further optimize the algorithm's implementation for a given target system.

7. Epilogue

7.1. Discussion

In our study of time-scaling methods we followed one particular line of thought towards computationally efficient high quality methods. We favored time-scaling based on time–frequency representations over model based approaches, and proceeded to present an iterative phase reconstruction method for time-scaled magnitude spec-

[9] As we used the same input signals and the same time-scaling factor for Figs. 6 and 2 and for Figs. 5 and 1, the reader might be interested in comparing these sets of figures to get an idea about the effectiveness of the proposed solution.

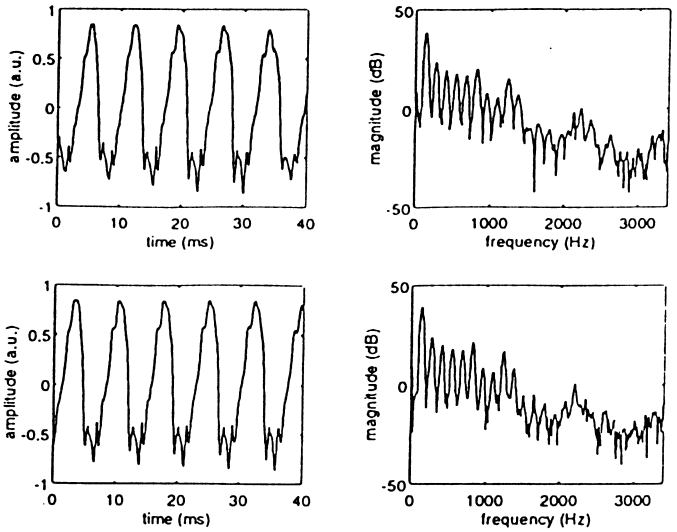

Fig. 5. Frequency domain effects of WSOLA time-scaling. The upper row shows a 40 ms voiced speech frame and its spectrum; the second row illustrates that when this signal is played at half speed using WSOLA no frequency shifting occurs.

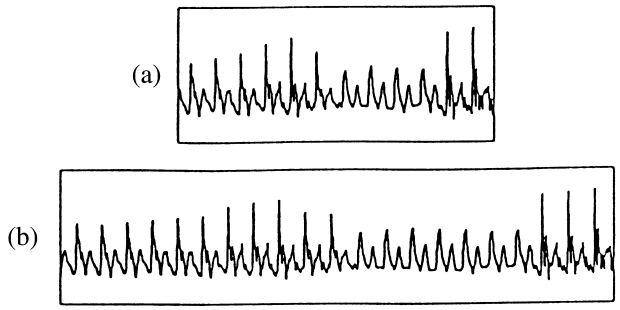

Fig. 6. Illustration of an original speech fragment (a) and the corresponding WSOLA output waveform when slowed down to 60% speed (b).

trograms. The search for a good initial phase estimate led us to consider synchronized overlap-add methods which were still further optimized to eventually arrive at WSOLA.

As a common feature of synchronized OLA algorithms we found that they operate in the time domain by performing a short-time analysis (i.e., a segmentation) and scaling only one of the two time dimensions. They avoid the need for actual frequency domain computations by synchronization of the segments that are used in OLA. Besides high computational efficiency, these methods provide high-quality time-scaling because they work with original segments of the input signal. These segments contain rich acoustic detail that would not be easily reproduced with purely model based approaches. Synchronized OLA methods are very flexible in that arbitrary time-warping functions $\tau(n)$ can be realized and time-scaling factors can be specified in a time-varying way if desired. Further, these methods are inherently robust since no assumption is made concerning the nature of the signal to be time-scaled (for example, no measures related to a speech production model are made and modified in this time-scaling).

As synchronized OLA algorithms are easily interpreted as automatic waveform editing methods (Weinrichter and Brazda, 1986) we can also easily see that they do have limitations to their capabilities. The structure of signals that can be

processed with good success must remain sufficiently simple (synchronization of segments can become a problem if the acoustic waveform is too complex (Spleesters et al., 1994). Fortunately this requirement is satisfied for speech. In fact, some useful results have even been shown in experiments on synchronized OLA time-scaling for digital audio signals (Spleesters et al., 1994; Laroche, 1993).

Although the quality of processed speech is indeed very high, a slight reverberance can occur when speech is slowed down by a significant amount. This too is easily explained using the waveform editing interpretation: some segments of the signal are repeated in order to slow down speech. A possible improvement could be obtained if the segments that are used more than once in the output signal are time reversed whenever they are repeated (Moulines and Charpentier, 1990). This helps to break up the repetitive structure and corresponds to sign inverting the phase, which is allowed in unvoiced speech but cannot be used for voiced segments.

7.2. Experimental evidence

Since more than a decade or so, OLA based techniques for prosodic modification have been studied in speech processing research labs around the world. The natural speech quality and the robustness of these methods, together with their ease of use, have reached such high levels that they are replacing traditional analysis–synthesis techniques in many areas. Current designs for concatenative text-to-speech synthesis, for instance, normally use OLA based techniques like PSOLA (Moulines and Charpentier, 1990; Valbret, 1992), MBROLA (Dutoit and Leich, 1992), PIOLA (Vogten et al., 1991), etc., for their prosodic modifications instead of traditional techniques like LPC (see, e.g., (Hess, 1992) for a discussion). These systems yield a segmental quality that is very close to human speech; their overall machine-like quality mostly depends on linguistic defects in computing a natural melody and rhythm for the input text (van Heuven and van Bezooijen, 1995).

Because the overall quality of OLA based speech modifications mostly depends on the quality of the specified prosodic contours themselves, it is difficult to formally evaluate the quality of these techniques per se. [10] During the last decade the author and his co-workers implemented several OLA based techniques and especially PSOLA, PIOLA and WSOLA in different styles on different platforms and systems, from Matlab over C to real-time Assembler code. In our experience these algorithms are very easy to implement and to operate and the resulting speech quality is very natural sounding and free from artifacts over a wide range of modification factors. Also, these algorithms have been applied with success in research projects (Spleesters et al., 1994; Verhelst, 1990; Verhelst and Borger, 1991; Verhelst, 1991, 1997) and real-life applications alike (Bellens and Passchyn, 1994).

7.3. Demonstration

A number of sound files [11] illustrate the capabilities of WSOLA time scaling. The original (unprocessed) recordings are typical low to medium quality sound files, recorded in an office environment on a contemporary multimedia PC. All processings were fully automatic and used the WSOLA algorithm described in Section 6.3 using the cross-correlation similarity measure with timing tolerance $\Delta_{\max} = 7$ ms. The segment length was 20 ms for the first set of demonstrations (the default value in our implementation) and 15 ms for the second set (they were produced one and a half years before the first set).

The first demonstration presents uniform time scalings of the original utterances COMING and STEP. The files named CxSy each contain a concatenation of COMING scaled to x% of its original speaking rate and STEP scaled to y% of its original speaking rate. They illustrate the typical

[10] Even if prosodic contours are interchanged between natural utterances of a same sentence, quality differences between natural and processed speech can often be ascribed to allophonic differences between the utterances forming less natural combinations of micro-prosodic and segmental features in the modified speech (Verhelst and Borger, 1991; Verhelst, 1997).

[11] See the speech files on www.elsevier.nl/locate/specom.

speech quality of WSOLA processing for male and female voices.

The second demonstration is a subset from an older demonstration concerning automatic post-synchronization of speech utterances and illustrates WSOLA processing in an application with time-varying time-scaling factors. As published in (Verhelst, 1997), the application uses dynamic time-warping to compute the time-relationship between two utterances of the same material and applies WSOLA processing with the so derived time-warping function to automatically synchronize one of the utterances with the other. The file MIX is a mix of two original utterances by different speakers. In LINWARP the utterances are mixed after uniform time-scaling and in DTWWARP they are mixed after non-uniform time-synchronization (also see Verhelst, 1997).

7.4. Application for frequency-scaling

In this section we present an application that illustrates the use of time-scaling methods for frequency-scaling. The application considered is a prototype system for playback of speech recordings from analog compact cassette at variable speed. The overall system is illustrated in Fig. 7.

In order to deliver speech at a desired output rate that can be different from the original recording rate, it is obvious that the tape speed must be varied accordingly and that the resulting signal will be frequency-scaled, as we discussed in Section 1. A time-scaling algorithm like WSOLA can be used to apply the inverse frequency-scaling to the output signal of the tape recorder to restore the original frequency domain characteristics of the recording.

If we let α denote the tape speed relative to the recording speed, the signal to be processed has a properly scaled duration $\Delta t/\alpha$ and a bandwidth $\alpha\Delta\omega$ where Δt is some time span measured in the original recording and $\Delta\omega$ is the original signal's bandwidth. As an analog low pass signal needs to be digitized at a frequency at least twice its bandwidth, the output of the tape recorder should be sampled at least at $f_s = 2\alpha\Delta\omega$.

Frequency-scale compensation can then be achieved in two steps. First the signal is time-scaled with the same factor α. This results in a signal whose bandwidth remains $\alpha\Delta\omega$ but where the original time span Δt would now last $\Delta t/\alpha^2$ if converted to analog with the same clock frequency f_s. By converting the time-scaled signal to analog with a DA convertor clocked at f_s/α, we restore its time span to $\Delta t/\alpha$ and its bandwidth to $\Delta\omega$.

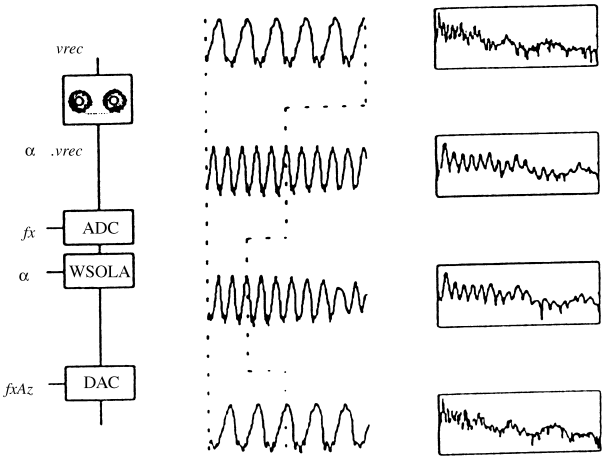

Fig. 7. Illustration of frequency shift compensation for analog recordings. The middle column shows the time domain representation of signals at the different positions of the block diagram in the left column; the right column shows the frequency domain representation of these signals (x-axis scales are 0–40 ms and 0–5 kHz, respectively; $\alpha = 2$).

Based on this strategy and using WSOLA, a prototype for 'speed reading' for the blind with a single DSP processor achieved real-time operation up to $\alpha = 2.5$ (Bellens and Passchyn, 1994).

The same strategy can also be used for bandwidth reduction during analog speech transmission. In that case, the speech signal is first time compressed by a factor α to obtain a signal of duration $\Delta t/\alpha$ and bandwidth $\Delta\omega$. Conversion to analog using a clock that runs α times slower than the AD conversion clock results in a signal of restored duration Δt with a bandwidth that is compressed by a factor α. After transmission, the original bandwidth can be restored by digital time-expansion and DA conversion using the original clock frequency.

References

Bellens, E., Passchyn, A., 1994. Frequentieshift compensatie bij weergave van spraakopnamen aan variabele snelheid. Unpublished MSEE thesis report.

d'Alessandro, C., 1989. Time-frequency modifications using an elementary waveform speech model. In: Proceedings of Eurospeech'89, pp. 211–214.

Deller Jr, J.R., Proakis, J.G., Hansen, J.H.L., 1993. Discrete-Time Processing of Speech Signals. MacMillan, New York.

Dutoit, T., Leich, H., 1992. Improving the TD-PSOLA Text-to-speech synthesizer with a specially designed MBE re-synthesis of the segments database. In: Vandewalle, J., et al. (Eds.), Signal Processing VI. Elsevier, Amsterdam, pp. 343–346.

Gabor, D., 1947. Acoustical quanta and the theory of hearing. Nature 159, 591–594.

Griffin, D.W., Lim, J.S., 1984. Signal estimation from modified short-time Fourier transforms. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (2), 236–248.

Hess, W.J., 1992. Speech synthesis – A solved problem? In: Vandewalle, J., et al. (Eds.), Signal Processing VI. Elsevier, Amsterdam, pp. 37–46.

Laroche, J., 1993. Autocorrelation method for high quality pitch/time scaling. In: Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, pp. 200–204.

Lindsay, P.H., Norman, D.A., 1977. Human Information Processing, Academic Press, New York.

Moore, B.C.J., 1982. An Introduction to the Psychology of Hearing, Academic Press, New York.

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

Moulines, E., Verhelst, W., 1995. Time-domain and frequency-domain techniques for prosodic modification of speech. In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, pp. 519–555.

Roads, C., 1998. The Computer Music Tutorial. MIT Press, Cambridge, MA.

Roucos, S., Wilgus, A., 1985. High quality time-scale modification of speech. In: Proceedings of ICASSP-85. IEEE, pp. 493–496.

Spleesters, G., Verhelst, W., Wahl, A., 1994. On the Application of automatic waveform editing for time warping digital and analog recordings. AES preprint 3843 (p 11.3) presented at the 96th Convention of the Audio Engineering Society, Amsterdam, 1994.

Valbret, H., 1992. Système de conversion de voix pour la synthèse de parole. ENST 92-E017.

van Heuven, V.J., van Bezooijen, R., 1995. Quality evaluation of synthesized speech. In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, pp. 707–738.

Verhelst, W., 1990. An implementation of the PSOLA/KDG Waveform synthesis technique. IPO Report 733. Institute for Perception Research.

Verhelst, W., 1991. On the quality of speech produced by impulse driven linear systems. In: Proceedings of ICASSP-91. IEEE, pp. 501–504.

Verhelst, W., 1997. Automatic postsynchronization of speech utterances. In: Proceedings of Eurospeech97. ESCA, pp. 899–902.

Verhelst, W., Borger, M., 1991. Intra-speaker transplantation of speech characteristics. In: Proceedings of Eurospeech91. ESCA, pp. 1319–1322.

Verhelst, W., Roelands, M., 1993. An overlap-add technique based on waveform similarity (WSOLA) for high-quality time-scale modification of speech. In: Proceedings of ICASSP-93. IEEE, pp. 554–557.

Vogten, L.L.M., Ma, C., Verhelst, W.D.E., Eggen, J.H., 1991. Pitch inflected overlap and add speech manipulation. European Patent 91202044.3 issued to Philips Eindhoven.

Weinrichter, H., Brazda, E., 1986. Time domain compression and expansion of speech. In: Young, I.T. (Ed.), Signal Processing. Elsevier, Amsterdam, pp. 485–488.