24. Voice Transformation

Y. Stylianou

Voice transformation refers to the various modifications one may apply to the sound produced by a person, speaking or singing. In this chapter we give a description of various ways in which one can modify a voice and provide details on how to implement these modifications using a simple, but quite efficient, parametric model based on a harmonic representation of speech. By discussing the quality issues of current voice transformation algorithms in conjunction with the properties of speech production and perception systems we try to pave the way for more-natural voice transformation algorithms in the future.

24.1 Background

Voice transformation refers to the various modifications one may apply to the sound produced by a person, speaking or singing. Voice transformation involves signal processing and the physics (or at least the understanding) of the speech production process and natural language processing. Driven mainly by its applications, signal processing has evolved faster than the physics of speech processing, even giving the impression that signal processing alone may be required to achieve high-quality voice transformation. To an external observer, this is similar to the problem of how to make an omelette without eggs. It is not surprising therefore that, although for some categories of voice transformation good quality of speech is produced, this is not true in general. While it is relatively easy to explain to a non-speech expert the necessity of speech modeling by providing examples from the history of telecommunications, this is not obvious for voice transformation. About two decades ago, it was easy to explain the applications of voice transformation technology to a speech processing engineer by providing examples from a specific area of speech technology: concatenating speech synthesis. In the late 2000s, providing a reason for voice transformation to both a speech expert and nonexpert faced the same difficulty. One reason for this was that the main application of voice transformation, that of concatenating speech synthesis, had evolved in a direction where it seemed that signal processing was no longer needed for this application. Such a point of view, however, was also supported by the quality problems perceived in a modified speech signal.

Recently, interest in voice transformation has increased substantially, and it is again the application of speech synthesis that is setting the pace. Voice

transformation is a flexible, possibly simple, and efficient way to produce the variety needed in the current text-to-speech (TTS) systems based on the concatenation of units (both large and small) [24.1]. In this chapter, we give a description of various ways in which one can modify a voice and provide details of how to implement these modifications using a simple, but quite efficient, parametric model based on a harmonic representation of speech. Discussing quality issues of current voice transformation algorithms in conjunction with properties of the speech production and perception systems we try to pave the way for more-natural voice transformation algorithms in the future.

24.2 Source–Filter Theory and Harmonic Models

24.2.1 Harmonic Model

When designing voice transformation techniques it is often convenient to refer to the source–filter model of speech production. According to this model, speech is viewed as the result of passing a glottal excitation signal (source) through a time-varying linear filter that models the resonant characteristics of the vocal tract. The most well-known source–filter system is that based on linear prediction (LP) of speech [24.2]. In its simplest form, a time-varying filter modeled as an autoregressive (AR) filter is excited by either quasiperiodic pulses (during voiced speech), or noise (during unvoiced speech). Many attempts have been made to improve the source (excitation) signal in the LP context. This includes multipulse LP [24.3], and code-excited linear prediction (CELP) [24.4]. A more-compact and at the same time flexible representation of the excitation signal has been proposed from a family of speech representations referred to as sinusoidal models (SM) [24.5]. In SM, the excitation signal for both voiced and unvoiced speech frames is represented by a sum of sinusoids:

$$e(t) = \sum_{k=0}^{K(t)} a_k(t) \mathrm{e}^{\mathrm{i}\phi_k(t)} \,, \tag{24.1}$$

where $a_k(t)$ and $\phi_k(t)$ are the instantaneous excitation amplitude and phase of the k-th sinusoid, respectively, and $K(t)$ is the number of sinusoids, which may vary in time. Especially for speech signals, a model where the sinusoids are harmonically related is quite valid (in the mean-squared-error (MSE) sense) while it allows a simple and convenient way of applying various modifications to the speech signal. In this case:

$$\dot{\phi}_k(t) = 2\pi k f_0(t) \,, \tag{24.2}$$

where $f_0(t)$ is the instantaneous fundamental frequency, which will also be referred to as the *pitch* in this chapter. Such a representation is still valid for both voiced and unvoiced speech frames. In the case of unvoiced speech frames a constant fundamental frequency is considered (i. e., 100 Hz) resulting in a Karhunen–Loeve representation of this speech category [24.5]. A further simplification of the excitation signal is convenient assuming that the excitation amplitude, $a_k(t)$, is constant over time and equal to unity: $a_k(t) = 1$. Based on these simplifications, the time-varying linear filter that models the resonant characteristics of the vocal tract approximates the combined effect of:

1. the transmission characteristics of the supraglottal cavities (including radiation at the mouth opening)
2. the glottal pulse shape

Its time-varying transfer function can be written

$$H(f;t) = G(f;t)\mathrm{e}^{\mathrm{i}\Psi(f;t)} \,, \tag{24.3}$$

where $G(f;t)$ and $\Psi(f;t)$ are, respectively, referred to as the time-varying amplitude and phase of the system. Speech processing is often (if not always) performed in a frame-by-frame basis, where each frame (i. e., about 20 ms) is considered to be a stationary process. In this case, inside a frame, the filter $H(f;t)$ is considered as linear time invariant (LTI). Then, the output speech signal $s(t)$ can be viewed as the convolution of the impulse response of the LTI filter, $h(t)$, and the excitation signal, $e(t)$:

$$s(t) = \int_0^t h(t-\tau)e(\tau)\mathrm{d}\tau \,. \tag{24.4}$$

Recognizing then that the excitation signal is just the sum of $K(t)$ eigenfunctions of the filter, $H(f)$, the

following speech model is obtained:

$$s(t) = \sum_{k=0}^{K(t)} G[f_k(t)]\, \mathrm{e}^{\mathrm{i}\{\phi_k(t)+\Psi[f_k(t)]\}} = \sum_{k=0}^{K(t)} A_k(t)\, \mathrm{e}^{\mathrm{i}\theta_k(t)}\,, \tag{24.5}$$

where $f_k(t) = k f_0(t)$ (the eigenfrequencies). The harmonic amplitude $A_k(t)$ of the k-th harmonic is the system amplitude $G[f_k(t)]$ (the eigenvalue). The phase $\theta_k(t)$ of the k-th harmonic is the sum of the excitation phase $\phi_k(t)$ and the system phase $\Psi[f_k(t)]$:

$$\theta_k(t) = \phi_k(t) + \Psi[f_k(t)]\,; \tag{24.6}$$

$\theta_k(t)$ is often referred to as the *instantaneous phase* of the k-th harmonic.

24.2.2 Analysis Based on the Harmonic Model

Parameters of the harmonic model of speech may be estimated by minimizing a least-squares criterion [24.6] or by minimizing a mean-squared error that leads to a simple *peak-picking* approach [24.5]. The peak-picking approach results in a sinusoidal rather than a harmonic model. A second step is then required to fit a harmonic model to this sinusoidal model by selecting the fundamental frequency that best represents the set of estimated sinusoids. At each analysis time instant, t_a^i, a set of parameters are estimated: the fundamental frequency, f_0^i, the harmonic amplitudes, A_k^i, and the harmonic phases, θ_k^i.

Use of the harmonic model for voice transformations is simplified if the distance between two successive analysis time instants is equal to the local pitch period, $P(t_\mathrm{a}^i) = 2\pi / f_0^i$:

$$t_\mathrm{a}^{i+1} = t_\mathrm{a}^i + P\left(t_\mathrm{a}^i\right)\,. \tag{24.7}$$

Another important step before synthesis is required: the estimation of the *amplitude and phase envelopes*, $A(f)$ and $\theta(f)$ (i. e., a continuous function of frequency) from the discrete set of amplitude and phase values, respectively.

While a number of methods can be used to estimate the amplitude envelope, for example, the linear prediction and homomorphic estimation techniques [24.7], it is desirable to use a method that yields an envelope that passes through the measured harmonic amplitudes. Such a technique was developed for the spectral envelope estimation vocoder (SEEVOC) [24.8] and was used in the sinusoidal model in [24.5]. Another approach was proposed in [24.9]: it provides a continuous frequency envelope when values of this envelope specified only at discrete frequencies (i. e., exactly the situation in the previously described harmonic representation). This approach makes use of cepstral coefficients and is based on a frequency-domain least-squares criterion combined with regularization to increase estimation robustness.

For the phase envelope $\theta(f)$, the previous techniques cannot be used since the phase values have been estimated modulo 2π (principal values). Therefore, a phase unwrapping algorithm has to be used. Two main approaches exist:

1. phase continuity by adding appropriate multiples of 2π to the principal phase values [24.10]
2. continuity by integration of the phase derivative

These algorithms try to obtain a continuous phase envelope in the frequency domain. An extension of these techniques to preserve the continuity in the time domain as well has been proposed using the information of phase from previous voiced frames [24.11].

An alternative to the phase envelope approach is the use of a minimum phase model for the system phase, while for the excitation phase a representation of the excitation in terms of its impulse locations (*onset times*) is used [24.12]. This approach, however, lacks robustness because estimation of the onset times requires precision that is not always easy to obtain.

Next, we will consider the case when a spectral envelope, $A(f)$, and phase envelope, $\theta(f)$, are provided.

24.2.3 Synthesis Based on the Harmonic Model

Without speech modification, synthesis time instants, t_s^i coincide with the analysis time instants t_a^i, i.e., $t_\mathrm{s}^i = t_\mathrm{a}^i$, $\forall i$.

Let $(A_k^i, \theta_k^i, f_0^i)$ and $(A_k^{i+1}, \theta_k^{i+1}, f_0^{i+1})$ denote the set of parameters at synthesis time instant t_s^i and t_s^{i+1} for the k-th harmonic, respectively. Amplitudes and phases are obtained by sampling the phase and amplitude (spectral) envelopes at the harmonics of the fundamental frequencies f_0^i and f_0^{i+1}. The instantaneous amplitude $A_k(t)$ is then obtained by linear interpolation of the

estimated amplitudes at the frame boundaries:

$$A_k(t) = A_k^i + \frac{A_k^{i+1} - A_k^{i+1}}{t_s^{i+1} - t_s^i} t \quad \text{for } t_s^i \le t < t_s^{i+1} . \tag{24.8}$$

In contrast to the third-order polynomial used in [24.13, 14], the harmonic model allows the use of a simple first-degree polynomial for the phase. First, the phase at t_s^{i+1} is predicted from the estimated phase at t_s^i by

$$\hat{\theta}_k^{i+1} = \theta_k^i + k2\pi f_{0\text{av}}\left(t_s^{i+1} - t_s^i\right) , \tag{24.9}$$

where $f_{0\text{av}}$ is the average value of the fundamental frequencies at t_s^i and t_s^{i+1}:

$$f_{0\text{av}} = \frac{f_0^i + f_0^{i+1}}{2} . \tag{24.10}$$

Next, the phase θ_k^{i+1} is augmented by the term $2\pi M_k$ (M_k is an integer) in order to approach the predicted value. Therefore, the value of M_k is given by

$$M_k = \left\langle \frac{1}{2\pi}\left(\hat{\theta}_k^{i+1} - \theta_k^{i+1}\right) \right\rangle . \tag{24.11}$$

Then, the instantaneous phase $\theta_k(t)$ is simply obtained by linear interpolation

$$\theta_k(t) = \theta_k^i + \frac{\theta_k^{i+1} + 2\pi M_k - \theta_k^i}{t_s^{i+1} - t_s^i} t , \quad t_s^i \le t < t_s^{i+1} . \tag{24.12}$$

Having determined the instantaneous values of the harmonic amplitudes and phases the estimated speech signal (a harmonic representation of the speech signal) is then obtained by:

$$\hat{s}(t) = \sum_{k=0}^{K} A_k(t) \cos\left[\theta_k(t)\right] , \tag{24.13}$$

where $A_k(t)$ is given by (24.8) and $\theta_k(t)$ by (24.12).

Based on the source–filter model various speech modification methods can now be defined. Some of these refer only to the source signal, others only to the filter, while others apply to both the source and filter. Moreover, by developing the source–filter model in the context of the harmonic representation of speech signals, a mathematical notation regarding these modifications can be introduced that will be used throughout the rest of the chapter.

24.3 Definitions

24.3.1 Source Modifications

Modifications in the source signal are usually referred to as *prosodic modifications* and include three main types: time-scale modification, pitch modification, and intensity modification.

Time-Scale Modification

The goal of time-scale modification is to change the apparent rate of articulation without affecting the perceptual quality of the original speech. This requires the pitch contour to be stretched or compressed in time, and the formant structure to be changed at a slower or faster rate than the rate of the input speech, but otherwise not be modified. Figure 24.1 shows an example of time stretching where the pitch period contour is slowed down but not modified.

Pitch Modification

The goal of pitch modification is to alter the fundamental frequency in order to compress or expand the spacing between the harmonic components in the spectrum while preserving the short-time spectral envelope (the locations and bandwidths of the formants) as well as the time evolution. In contrast to time-scale modifica-

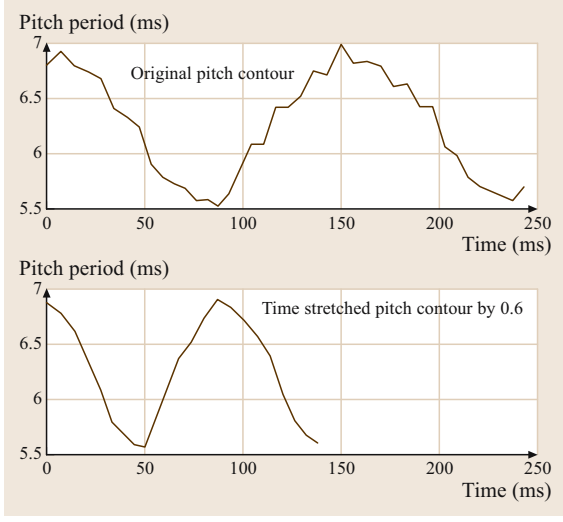

Fig. 24.1 Pitch-period contour: original and time-stretched by a factor of 0.6

tions, in this case the pitch contour is modified *without* modifying, however, the time resolution of the pitch contour. Figure 24.2 shows an example of pitch modification by constant pitch-scale factor (0.6): the time evolution is preserved and the pitch-period contour is scaled by 0.6. In this case, the input fundamental frequency is increased by a factor of $1/0.6$. This could give the impression that a male voice will sound more like a female voice, while a female voice will more sound like a child's voice.

Intensity Modification

It is widely considered that intensity modification is the simplest modification among the prosodic modifications. This is because it can be easily performed by associating an intensity scale factor at each analysis time instant of a signal. The signal is then just multiplied by this scale factor. In the case of a parametric model like the harmonic model developed previously, the scale factor is applied to the harmonic amplitudes $A_k(t)$ in (24.5). It may seem strange to modify a prosodic feature by changing a parameter corresponding to the vocal-tract filter. However, it should be remembered that the filter has been considered as an LTI filter; therefore, multiplying the amplitude of the excitation signal by a constant results in multiplying the harmonic amplitudes, $A_k(t)$, by the constant.

24.3.2 Filter Modifications

By filter modification we mean the modification of the magnitude spectrum of the frequency response of the vocal tract system, $|H(\omega)|$. It is widely accepted that $|H(\omega)|$ carries information of speaker individuality. Representations of the magnitude spectrum (i. e., mel frequency cepstral coefficients (MFCC), line spectrum frequencies (LSF), etc.) have been used a lot in the area of speaker identification and recognition as well as for speaker normalization for robust speech recognition. Therefore, by modifying the magnitude spectrum of the vocal tract, speaker identity may be controlled. We may distinguish two types of filter modification, which are described below.

Without a Specific Target

In this case, the filter characteristics of a speaker are modified in a general way without having a specific target (speaker) in mind. For example, we may wish to modify the overall quality of a speech signal produced by a female speaker so that it sounds as if it had been produced by an older female speaker. Based on the source–filter theory for the production of speech

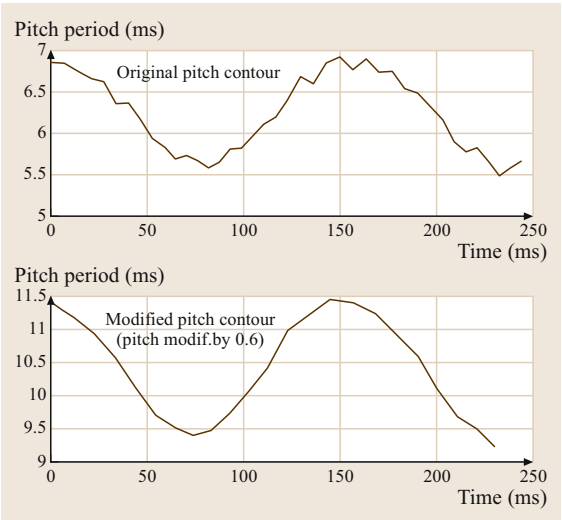

Fig. 24.2 Pitch-period contour: original and modified by applying a pitch modification factor of 0.6

we know that the formants for a female voice are distributed in higher frequencies than the formants from an older female voice. Similar observations are valid for the harmonic frequencies. Therefore, one can modify in a general way the power spectrum of a speaker so that the resulting spectrum has the characteristics of a family of speakers (child's, old person's voices, etc.). Figure 24.3 shows an example of filter modification to modify the spectrum from a female voice to a spectrum *similar* to an old female person. For this one should compress the

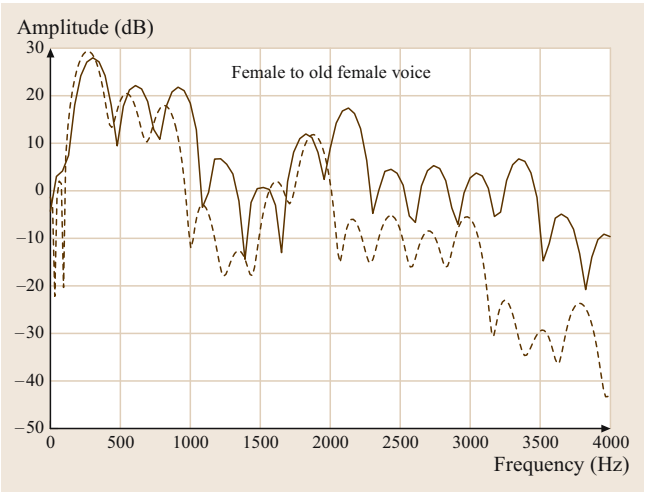

Fig. 24.3 Female (*solid line*) to an old female (*dashed line*) filter modification

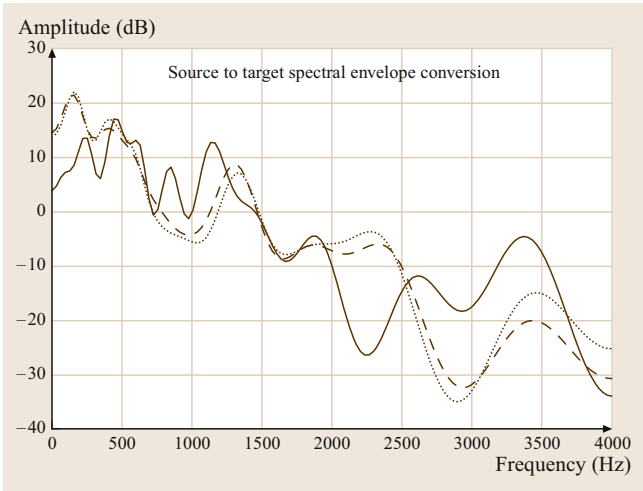

Fig. 24.4 Source (*solid line*) to target (*dashed line*) filter transformation. The transformed filter is shown by a *dashed–dot line*

frequency information so that formants and harmonics are moved towards lower frequencies. The result shown in Fig. 24.3 was obtained by combining two operations: lowering the sampling frequency of the source signal (female voice, 44 100 Hz) to 18 000 Hz, and then applying time-scale modification with a factor of 8/9. The result signal can be played back at 16 000 Hz without any modification in the articulation rhythm. The quality of the modified signal is similar to that the initial signal.

With a Specific Target

In this case, the filter of a speaker (the source speaker) is modified in such a way that the modified filter approximates in the mean-squared sense the characteristics of the filter of another speaker (the target speaker). Usually we refer to this type of modification as a *transformation* or *conversion*. An example of a such transformation is depicted in Fig. 24.4. In this example the original spectrum is shown by a solid line, the target spectrum by a dashed line, and the transformed original to target spectrum by a dashed–dot line. To obtain the transformed spectrum, there is a learning process using many similar examples of the source and the target spectrum; therefore, the transformed spectrum is equal to the average spectrum of the target spectra used during the training process. Details about this type of spectral transformation will be provided in the next section.

24.3.3 Combining Source and Filter Modifications

To transform a female into a male voice, performing filter modification or only pitch modification alone may not provide convincing results. In most cases, source and filter modifications must be combined. For example, the prosody characteristics of a speaker may be a critical cue used for the identification of the speaker by others (speaking style) while at the same time, vocal-tract characteristics are also important for identification. Therefore, if we want to modify the voice of the speaker so that it sounds like the voice of another speaker, prosody and vocal tract modifications should be combined. If a target speaker is provided then this combined source and filter modifications is referred to as *voice conversion* or *transformation*. In contrast, when a specific target is not provided, this is usually referred to as *voice modification*.

Voice morphing is another type of combined source and/or filter modifications. In this case the same two sentences are uttered by two speakers and then a third speaker may be generated having characteristics from both speakers. This is mainly achieved by a dynamic time warping (DTW) algorithm between the two sentences, aligning the acoustic data and then applying a linear or other type of interpolation between the aligned data (source and/or filter characteristics). Sometimes this type of voice transformation is confused with voice conversion. Note that in voice conversion the sentence to be converted, uttered by the source speaker, *has never been uttered* by the target speaker. However, in voice morphing there are two source speakers that generate a new voice saying the same text as the two source voices. In voice conversion, there is only one source speaker and one target speaker, and the voice characteristics of the source speaker should be transformed into the voice characteristics of the target speaker (i. e., a new speaker is not generated in this case).

In the next section we will provide details about the main prosodic (time and pitch) modifications and the filter modifications. Then a system for voice conversion will be presented.

24.4 Source Modifications

Pitch synchronous analysis is the key to the simplicity of many source (prosodic) modification algorithms and is defined as follows. Given an analysis time instant, t_a^i, the next analysis time instant, t_a^{i+1} is determined by the

local pitch period at t_a^i, $P(t_a^i)$, using (24.7). The length of the analysis window is proportional to the local pitch period (usually two local pitch periods are used). We may distinguish two types of pitch synchronous analysis. The first one may be referred to as *strict* pitch synchronous analysis, where the analysis time instants are supposed to coincide with the glottal closure instants (GCIs, sometimes called *pitch marks*). In the other, referred to as *relaxed* pitch synchronous analysis, the analysis time instants do not (necessarily) coincide with the GCIs. Since the estimation of pitch marks from the speech signal is not a robust process, resulting sometimes in an incoherent synthesis, relaxed pitch synchronous analysis seems to be easier to use than the fixed approach. However, this is not true. Pitch modification requires the re-estimation of phases. Coherent synthesis mainly means synthesis without linear phase mismatches. Strict pitch synchronous methods explicitly remove any linear phase mismatch between successive frames by using GCIs. Relaxed pitch synchronous methods, however, need to re-estimate the linear phase component for the new pitch values, which is not a trivial task. Phase models [24.12] and estimation of phase envelopes [24.11] try to overcome these problems.

For the system presented here, we will consider that the analysis time instants (Sect. 24.2.2) have been determined in a relaxed pitch synchronous way.

Next, we will see how the pitch synchronous scheme allows the use of simple and flexible techniques for time-scale and pitch-scale modifications. The first step consists of finding out the synthesis time instants t_s^i (or synthesis pitch marks) according to the desired time-scale and pitch-scale modification factors. The modified signal is then obtained by using the new synthesis time instants.

24.4.1 Time-Scale Modification

We recall that the objective of time-scale modification is to alter the apparent rate of articulation without affecting the spectral content: the pitch contour and time evolution of the formant structure should be time scaled, but otherwise not modified [24.15].

From the stream of analysis time instants t_a^i and the desired time-scale modification factor $\beta(t)$, $(\beta(t) > 0)$ the synthesis time instants t_s^i will be determined. The mapping $t_a^i \rightarrow t_s^i = D(t)$ is referred to as the time-scale warping function, which is defined as the integral of $\beta(t)$:

$$D(t) = \int_0^t \beta(\tau)\,\mathrm{d}\tau \ . \tag{24.14}$$

Note that for a constant time modification rate $\beta(t) = \beta$, the time-scale warping function is linear: $D(t) = \beta t$. The case $\beta > 1$ corresponds to slowing down the rate of articulation by means of a time-scale expansion, while the case $\beta < 1$ corresponds to speeding up the rate of articulation by means of a time-scale compression. Thus, speech events that take place at a time t_{or} in the original time scale will occur at a time $t_{mo} = \beta t_{or}$ in the new (modified) time scale.

As an example, let us assume that at each analysis tim -instant t_a^i a time-scale modification factor b_s has been specified. Thus, $\beta(t)$ is a piecewise constant function, i. e., $\beta(t) = \beta_s$, $t_a^i \le t < t_a^{i+1}$. It follows therefore that the time-scale warping function $D(t)$ can then be written

$$D(t) = D(t_a^i) + \beta_s\left(t - t_a^i\right) \ , \quad t_a^i \le t < t_a^{i+1} \tag{24.15}$$

with $D(t_a^1) = 0$.

Having specified the time-scale warping function $D(t)$, the next step consists of generating the stream of the synthesis time instants t_s^i, *while preserving the pitch contour*: the pitch in the time-scaled signal at time t should be as close as possible to the pitch in the original signal at time $D^{-1}(t)$. In other words, $t \rightarrow P'(t) = P[D^{-1}(t)]$. We now have to find a stream of synthesis pitch marks (synthesis time instants) t_s^i, such that $t_s^{i+1} = t_s^i + P'(t_s^i)$. To solve this problem, the use of a stream of *virtual pitch marks*, t_v^i, in the original signal related to the synthesis pitch-marks by

$$\begin{aligned} t_s^i &= D\left(t_v^i\right) \ , \\ t_v^i &= D^{-1}\left(t_s^i\right) \ , \end{aligned} \tag{24.16}$$

is proposed in [24.15]. Assuming that t_s^i and t_v^i are known, we determine t_s^{i+1} (and t_v^{i+1}), such that $t_s^{i+1} - t_s^i$ is approximately equal to the pitch in the original signal at time t_v^i. This can be expressed as

$$t_s^{i+1} - t_s^i = \frac{1}{t_v^{i+1} - t_v^i} \int_{t_v^i}^{t_v^{i+1}} P(t)\,\mathrm{d}t \tag{24.17}$$

with $t_s^{i+1} = D(t_v^{i+1})$. According to this equation, the synthesis pitch period $t_s^{i+1} - t_s^i$ at time t_s^i is equal to the mean value of the pitch in the original signal calculated over the time interval $t_v^{i+1} - t_v^i$. Note that this time interval $t_v^{i+1} - t_v^i$ is mapped to $t_s^{i+1} - t_s^i$ by the mapping function $D(t)$.

The integral equation (24.17) is easily solved because $D(t)$ and $P(t)$ are piecewise linear functions. Figure 24.5 illustrates an example of the computation of synthesis pitch marks for time-scale modification by 1.5.

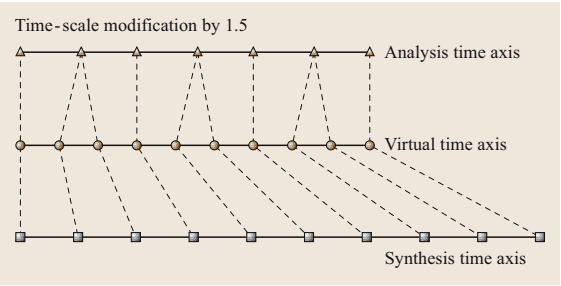

Fig. 24.5 Computation of the synthesis pitch marks for time-scale modification by 1.5

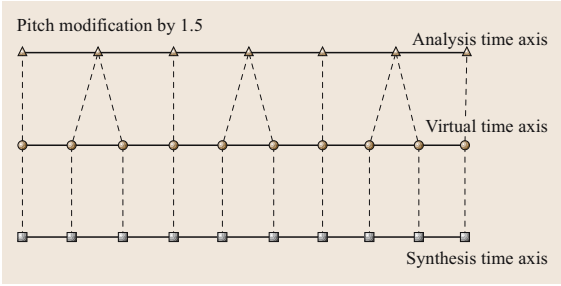

Fig. 24.6 Computation of the synthesis pitch marks for pitch modification by 1.5

24.4.2 Pitch Modification

The goal of the pitch-scale modification is to alter the fundamental frequency of a speaker while the spectral envelope of the speaker's vocal-tract system function is unchanged. Obviously, pitch modification is only applied on the voiced speech frames. The first step consists of computing the synthesis time instants t_s^i from the stream of the analysis time instants t_a^i and the pitch-scale modification factors $a(t)$, with $a(t) > 0$. We recall that the analysis time instants are set in a pitch synchronous way. We require the same for the synthesis time instants: $t_s^{i+1} = t_s^i + P'(t_s^i)$, where $P'(t_s^i)$ is approximately equal to the pitch period in the original signal around time t_a^i scaled by $1/\alpha(t_a^i)$:

$$P'\left(t_s^i\right) = \frac{P\left(t_a^i\right)}{\alpha\left(t_a^i\right)} \,. \tag{24.18}$$

Given the synthesis time instant t_s^i, the next synthesis time instant t_s^{i+1} is obtained by setting the synthesis pitch period to be equal to the mean value of the scaled pitch period (by $1/\alpha(t_a^i)$) in the original signal calculated over the time frame $t_s^{i+1} - t_s^i$:

$$t_s^{i+1} - t_s^i = \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s^i}^{t_s^{i+1}} \frac{P(t)}{\alpha(t)} \mathrm{d}t \,. \tag{24.19}$$

This integral equation is easily solved as $P(t)$ is a piecewise linear function and $\alpha(t)$ is a piecewise constant function:

$$\alpha(t) = a\left(t_a^i\right) \quad \text{for } t_a^i \le t < t_a^{i+1} \,. \tag{24.20}$$

Figure 24.6 shows an example of a mapping between the analysis and synthesis time instants for pitch modification by 1.5.

24.4.3 Joint Pitch and Time-Scale Modification

Based on the procedures presented above, joint pitch and time-scale modifications can easily be obtained. Given a pitch and time-scale modification factor at each analysis time instant and combining (24.17) and (24.19), the synthesis time instants can be obtained by solving the following integral equation

$$t_s^{i+1} - t_s^i = \frac{1}{t_v^{i+1} - t_v^i} \int_{t_v^i}^{t_v^{i+1}} \frac{P(t)}{\alpha(t)} \mathrm{d}t \,. \tag{24.21}$$

The synthesis time instants determined by the procedures above are not, in general, univocally associated with the analysis time instants (see for example Figs. 24.5 and 24.6). A solution consists of replacing the virtual pitch marks by the nearest analysis time instant. In this case, however, another problem arises: two or more successive frames could be the same and then the harmonic parameters (amplitudes and phases) will not vary within the frame, meaning that a high-quality modified synthetic signal is not produced. It follows therefore that the repeated analysis time instants should be elimi-

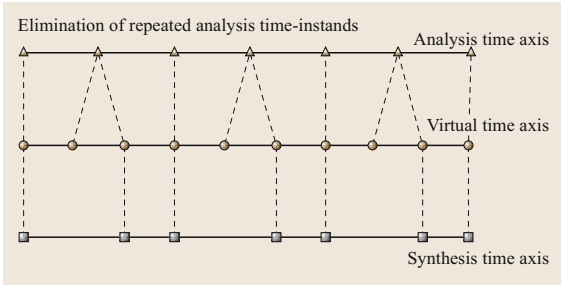

Fig. 24.7 Elimination of repeated analysis time instants from the example in Fig. 24.6

nated. As an example, in Fig. 24.7 the repeated analysis time instants in Fig. 24.6 are eliminated.

24.4.4 Energy Modification

Energy modification is performed by associating an intensity scale factor, $c(t_s^i)$, at each synthesis time instant. Then, the harmonic amplitudes are multiplied by the square root of the current harmonic intensity scale factor

$$A'_k(t_s^i) = \sqrt{c(t_s^i)}\,A_k(t_s^i) \quad \text{for } k = 1, \ldots, K\,, \tag{24.22}$$

where K is the number of harmonics. The modified amplitudes should then be used in estimating the amplitude envelope (Sect. 24.2.2).

24.4.5 Generating the Source Modified Speech Signal

In synthesis, the first step is the computation of the harmonic amplitudes and phases at the shifted harmonic frequencies. Note that in the case of time-scale modifications, the original amplitudes and phase may be preserved. In general, therefore, the amplitudes and phases are obtained by sampling the phase and amplitude envelopes at the corresponding harmonic frequencies. In the case of pitch modification, the phase and amplitude envelope should be sampled using the modified fundamental frequency: $f'_0(t_a^i) = \alpha(t_a^i) f_0(t_a^i)$. Given a spectrum, this results in a different number of harmonics from those initially included in the spectrum. When $\alpha(t_a^i) > 1$, fewer harmonics are included in the spectrum, and when $\alpha(t_a^i) < 1$, more harmonics are included. This means that the initial energy of the signal will be changed. Therefore, the amplitudes of the shifted harmonics are normalized in such a way that the final energy of the pitch-modified signal is equal to the energy of the unmodified one.

Using the new synthesis time instants and the modified set of parameters (amplitudes and phases) corresponding to each time instant, the source modified speech signal is obtained in exactly the same way as shown in Sect. 24.2.3. Examples of time-scale modification and pitch modification, both using a modification factor of 1.3, are depicted in Figs. 24.8 and 24.9, respectively.

An alternative to the parametric harmonic model for source modifications is the time-domain pitch synchronous overlap add (TD-PSOLA) [24.16]. TD-PSOLA relies heavily on the source–filter speech production model, although the parameters of this model are not estimated explicitly. TD-PSOLA is characterized by simplicity and low computational complexity, allowing good-quality prosodic modifications of speech. It is widely adopted for text-to-speech synthesis based on concatenation of acoustic units like diphones. Other methods similar to TD-PSOLA have also been proposed, including multiband resynthesis

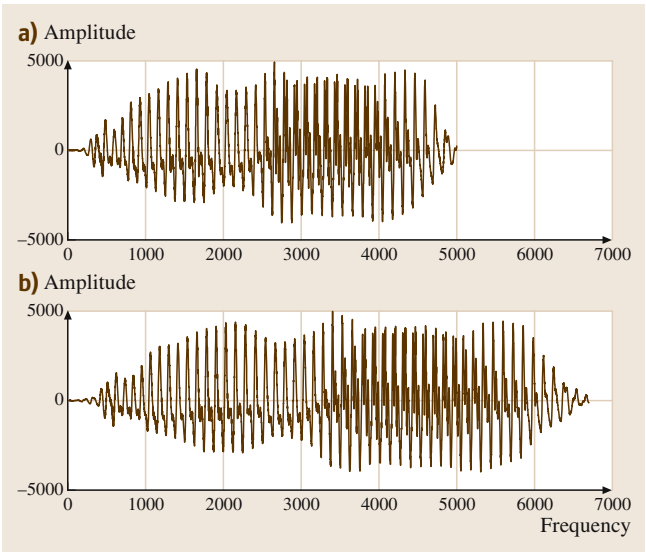

Fig. 24.8 (a) Original speech signal. (b) Time-scaled by 1.3. Time is provided in samples (sampling frequency: 16 kHz)

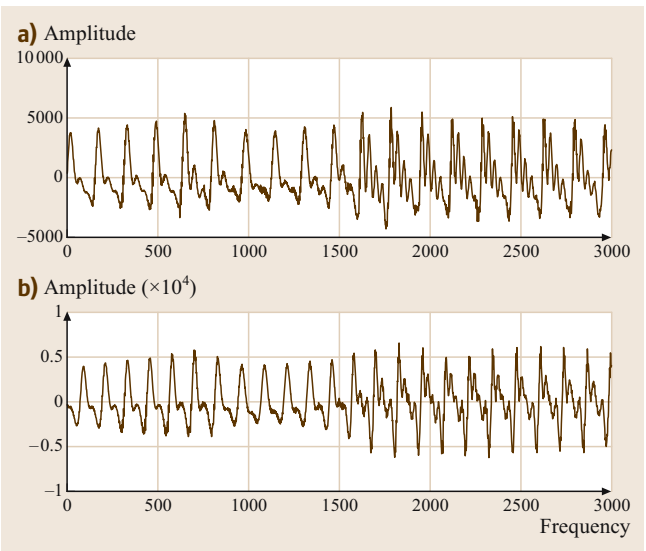

Fig. 24.9 (a) Original speech signal. (b) Pitch modified by 1.3. Time is provided in samples (sampling frequency: 16 kHz)

overlap add (MBROLA) [24.1]. The synchronized overlap add (SOLA) [24.17] and the waveform similarity overlap add (WSOLA) [24.18] methods have mainly been proposed for time-scale modification. In SOLA and WSOLA, successive speech frames to be overlapped are cross-correlated, providing the time shift to ensure that the two overlapping frames are synchronized and thus add coherently. TD-PSOLA relies on GCIs (pitch marks for voice sounds) to synchronize the speech frames. The use of pitch marks allows TD-PSOLA to apply a simple mechanism for pitch modification; this is not possible for the SOLA and WSOLA techniques. Examples of how to apply TD-PSOLA for speech modifications using the mapping of analysis and synthesis time instants are provided in Chap. 19.

Nonparametric approaches such as TD-PSOLA, SOLA, and WSOLA, do not allow complex modifications of the signal, such as increasing the degree of friction, or changing the amplitude and phase relationships between the pitch harmonics. Another major drawback is the manipulation of *noise-like* sounds presented in speech. For example, TD-PSOLA eliminates/duplicates short-time waveforms extracted from the original speech signal by windowing. When this approach is applied to unvoiced fricatives a tonal noise is produced because the repetition of segments of a *noise-like* signal produces an artificial long-time autocorrelation in the output signal, perceived as some sort of periodicity [24.15]. A simple solution to this problem consists of reversing the time-axis whenever the TD-PSOLA algorithm needs to duplicate *unvoiced* short-time signals [24.15]. This solution *reduces* the undesirable correlation in the output signal but the tonal quality does not completely disappear. This solution cannot be applied when the time-scale factor is greater than 2. Moreover, this solution cannot be used when voiced fricative frames are processed.

On the other hand, sinusoidal models have been found to be an efficient representation of voiced speech [24.5]. For a flexible representation of the unvoiced sounds and for high-quality pitch and time-scale modification of speech, hybrid [i. e., harmonic plus noise (HNM) [24.6]] models are more suitable.

24.5 Filter Modifications

Next, we will consider the case of filter modification *with* a specific target. This provides a more-general framework for filter modification than without a specific target, since it has a higher time resolution; the modification filter should be changed faster than in the case where a nonspecific target is provided. In this context, a set of source and target spectral envelopes is assumed given that an appropriate representation of the vocal tract spectral envelope is provided (i. e., using cepstral coefficients, line spectrum frequencies, mel frequency cepstral coefficients). To convert the source spectral envelope to the target spectral envelope, a training (or learning step) is necessary. During this step, a conversion function is trained. For this purpose, the source and the target speaker utter the same sentences.

One of the earliest approaches to the filter conversion is the mapping codebook method [vector quantization (VQ)] of *Abe* et al. [24.19], which was originally introduced for speaker adaptation by *Shikano* et al. [24.20]. The basic idea of this technique is to make mapping codebooks that represent the correspondence between the two speakers. A conversion of acoustic features from one speaker to another is therefore reduced to the problem of mapping the codebooks of the two speakers [24.19]. The main shortcoming of this method is the fact that the acoustic space of the converted signal is limited to a discrete set of envelopes. To avoid the limitations of the discrete space represented by VQ, a fuzzy vector quantization (FVQ) has been proposed by *Kuwabara* et al. [24.21]. A quite different approach, also based on VQ, has been proposed by *Iwahashi* et al. [24.22] using speaker interpolation. The use of linear multivariate regression (LMR) for *mapping* one class from the VQ space of the source speaker to the *corresponding* class in the VQ space of the target speaker has been proposed by *Valbret* et al. [24.23]. In the same communication [24.23], *Valbret* et al. proposed a spectral transformation approach based on dynamic frequency warping (DFW). In LMR a simple linear transformation function for each class has been proposed, while in DFW a third-order polynomial is used. All these methods have been developed in the context of VQ. Most authors agree that the mapping codebook approach, although it provides an impressive perceptive voice conversion effect, is plagued by poor quality and lack of robustness [24.24]. Approaches based on LMR and DFW also introduce discontinuities into the spectral information, as the acoustic space of a speaker

is partitioned in discrete regions. DFW succeeds in moving the formant frequencies but it has little or no effect on their amplitudes and bandwidths. Mapping functions have also been proposed using more-robust modeling compared to VQ of the acoustic space of a speaker based on the Gaussian mixture model (GMM). Assuming that the source and target vectors obtained from the speakers acoustic space are jointly Gaussian, a continuous probabilistic mapping function based on GMM has been proposed [24.25, 26]. A similar mapping function has been proposed by *Kain* et al. [24.27], jointly modeling the source and target vectors with GMM.

All of these techniques are based on parallel training data, where both the source and target speaker utter the same sentence. Then, a DTW is used to align the two signals in time in order to extract the aligned source and target training vectors. Approaches without the requirement of parallel data have also been proposed in the literature [24.28, 29]. However, using the same mapping functions for parallel and nonparallel data, it has been shown that training with parallel data provides better conversion results [24.28].

In the following, we will present a state-of-the-art system for filter modification (spectral conversion) based on GMM making use of parallel data [24.26].

24.5.1 The Gaussian Mixture Model

The Gaussian mixture density is a weighted sum of m component densities and given by the equation

$$p(\boldsymbol{x}|\Theta) = \sum_{i=1}^{m} \alpha_i \, p_i(\boldsymbol{x}|\theta_i) \,, \tag{24.23}$$

where $\boldsymbol{x} = [x_1 \; x_2 \; x_3 \; \cdots \; x_p]^{\mathrm{T}}$ is a p-by-1-dimensional random vector, $p_i(\boldsymbol{x}|\theta_i)$, for $i = 1, \cdots, m$, are the component densities and α_i are the mixture weights. Each component density, $p_i(\boldsymbol{x}|\theta_i)$, is a p-dimensional normal distribution

$$p_i(\boldsymbol{x}|\theta_i) = N(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{24.24}$$

with $\boldsymbol{\mu}_i$ the p-by-1 mean vector and $\boldsymbol{\Sigma}_i$ the p-by-p covariance matrix. The mixture weights, α_i, are normalized positive scalar weights ($\sum_{i=1}^{m} \alpha_i = 1$ and $\alpha_i \geq 0$). This ensures that the mixture is a true probability density function (PDF). The complete Gaussian mixture density is parameterized by the mixture weights, the mean vectors and the covariance matrices from all component densities, which is represented by the notation,

$$\Theta = (\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \,, \quad i = 1, \cdots, m \,. \tag{24.25}$$

The Gaussian mixture model (GMM) is a classic parametric model used in many pattern-recognition techniques [24.30] and speech applications such as speaker recognition [24.31]. In the GMM context, a speaker's voice is characterized by m acoustic classes representing some broad phonetic events, such as vowels, nasal or fricatives. The probabilistic modeling of an acoustic class is important since there is variability in features coming from the same class due to variations in pronunciation and co-articulation. Thus, the mean vector $\boldsymbol{\mu}_i$ represents the average features for the acoustic class ω_i, and the covariance matrix $\boldsymbol{\Sigma}_i$ models the variability of features within the acoustic class.

GMM parameters are usually estimated by a standard iterative parameter estimation procedure, which is a special case of the *expectation-maximization* (EM) algorithm [24.32, 33]. Initialization of the algorithm may be provided by VQ.

24.6 Conversion Functions

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be a set of p-dimensional vectors corresponding to the spectral envelopes of the source and the target speaker, respectively, where $t = 1, \ldots, n$. It is therefore assumed that these vectors have been obtained by speech samples from both speakers that have been aligned in time using a classic dynamic time-alignment algorithm [24.34]. We also assume that a Gaussian mixture model ($\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$, for $i = 1, \ldots, m$) has been fitted to the source vectors ($\boldsymbol{x}_t, t = 1, \ldots, n$).

It is worth noting that, if we take the limit case where the GMM is reduced to a single class and if the source vectors $\boldsymbol{x}_t$ follow a Gaussian distribution $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that the source and target vectors are jointly Gaussian, the minimum mean-square error estimate of the target vector is given by [24.35]

$$E[\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{x}_t] = \boldsymbol{\nu} + \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}) \,, \tag{24.26}$$

where $E[]$ denotes expectation, and $\boldsymbol{\nu}$ and $\boldsymbol{\Gamma}$ are, respectively, the mean target vector

$$\boldsymbol{\nu} = E[\boldsymbol{y}] \,,$$

and the cross-covariance matrix of the source and target vectors

$$\boldsymbol{\Gamma} = E[(\boldsymbol{y} - \boldsymbol{v})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}] ,$$

where the superscript T denotes transposition [24.36].

In [24.27], a direct extension of (24.26) to the GMM case was proposed. However, such an extension is not supported by the theory of statistics. Moreover, such an extension makes the assumption of a one-to-one correspondence between the source and target space, which is not valid in practice. To overcome these difficulties, and motivated by (24.26) the following conversion function between the source and the target data has been proposed [24.25]:

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{i=1}^{m} P(\omega_i|\boldsymbol{x}_t)\left[\boldsymbol{v}_i + \boldsymbol{\Gamma}_i \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_i)\right] . \tag{24.27}$$

The conversion function $\mathcal{F}$ is entirely defined by the p-dimensional vectors $\boldsymbol{v}_i$ and the p-by-p matrices $\boldsymbol{\Gamma}_i$, for $i = 1, \ldots, m$ (where m is the number of mixture components). This means that $\boldsymbol{v}_i$ and $\boldsymbol{\Gamma}_i$ are the parameters to be estimated. The parameters of the conversion function are computed by least squares optimization on the learning data so as to minimize the total squared conversion error

$$\epsilon = \sum_{t=1}^{n} ||\boldsymbol{y}_t - \mathcal{F}(\boldsymbol{x}_t)||^2 . \tag{24.28}$$

Since the conversion function given by (24.27) is linear, the optimization of its parameters is equivalent to the resolution of a set of linear equations in the least-squares sense. Details of the minimization of (24.28) may be found in [24.37]. The mapping function (24.27) can be used with full or diagonal covariance matrices. Note that the conversion function is reduced to

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{i=1}^{m} P(\omega_i|\boldsymbol{x}_t)\boldsymbol{v}_i \tag{24.29}$$

if the correction term that depends on the difference between the source vector $\boldsymbol{x}_t$ and the mean of the GMM component $\boldsymbol{\mu}_i$ in (24.27) is omitted. This reduced conversion function is similar to the formula proposed by *Abe* et al. [24.19] in the mapping codebook approach. Comparing (24.29) and (24.27) it follows that in VQ-type mapping functions the variability of the transformed spectral envelope is strongly restricted.

An example of filter conversion based on the approach described in this section has already been presented in Fig. 24.4.

24.7 Voice Conversion

Combining source and filter modifications a system that controls speaker's quality and individuality can be obtained. Continuing with the harmonic model as an example of representing speech, a speech signal produced by the source speaker is analyzed in a pitch-synchronous way (Sect. 24.2.1) to extract a set of source spectral envelopes. Given the spectral conversion function in (24.27), the target spectral envelopes in each frame are estimated by applying (24.27) to the source spectral envelopes. The next step is to apply the appropriate prosodic or source modifications in order to capture the *prosodic patterns* of the target speaker (Sect. 24.4). For this, prosodic profile analysis (i. e., characteristic articulation rate, stress, and emotions, pitch fluctuations) is required for both speakers. Determining such a profile is not a trivial task and has not yet been achieved in a convincing way. Thus, most voice conversion systems today make use

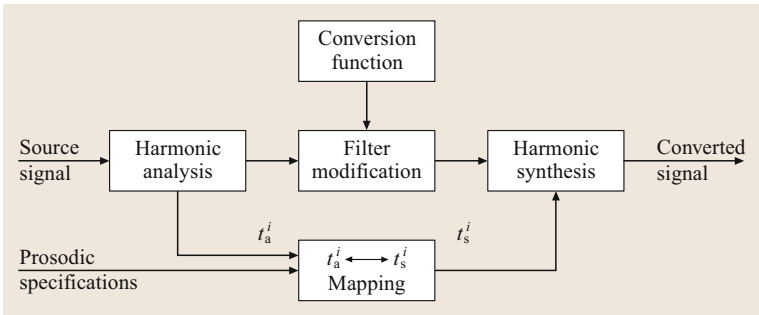

Fig. 24.10 Block diagram of a voice conversion system; t_a^i and t_s^i represent the analysis and synthesis time instants, respectively

of average prosodic modification factors. A block diagram for the conversion system based on the harmonic model and the voice conversion function is presented in Fig. 24.10.

24.8 Quality Issues in Voice Transformations

Voice transformations are usually evaluated in subjective tests. The overall impression from the results obtained from these tests is that time-scale modification is quite successful for moderate scale factors, while pitch-modified signals by pitch scale factors over 1.5 and below 0.7, suffer from various artifacts, making listeners classify the modifications as not natural. Voice conversion reaches a high score for transforming the identity of the source speaker to that of the target speaker. However, there are serious quality problems, which are mostly referred to as *muffling effects*. To improve the quality of the speech produced by various proposed voice transformation algorithms a better understanding of speech production and perception mechanisms is necessary. For example, when we want to increase the loudness of our voice while sitting in a cafeteria, we add stress to a *part* of our speech signal, like consonants, and *not* to all the speech events we produce. One hypothesis for this is that consonants carry more of the information load, which is connected with the intelligibility of the message we would like to transmit. According to this hypothesis we only increase the stress to these sounds by an amount that is sufficient to mask the cafeteria noise. Increasing the stress does not mean that the amplitudes of all the frequencies for this sound are increased. Stress means an increase of the subglottal pressure, which will result in an abrupt glottal closure by accentuating the Bernoulli effect on airflow through the glottis [24.38]. This corresponds to more energy *mostly* at high frequencies. From this example, it is obvious that even a simple intensity modification is not as simple as we thought. Continuing the above example, the increase of the subglottal pressure will increase the tension in the vocal folds, resulting in an increase of the pitch. This shows that modifying one parameter may require the modification of another as well.

In most Western languages consonants (we recall that consonants carry important information load) are shorter in duration than vowels (which carry more-prosodic information). Our perceptual system requires some time to process the perceived sounds. When we want to speak faster, we somehow *protect* the consonants. *Pickett* [24.39] has done extensive studies on the degree of change in vowels and consonants in speaking at a faster or slower rate. In [24.39], it was reported that, when going from normal to the faster rate, the vowels were compressed by 50% while the consonants were compressed by 26%. However, going from the slowest to the fastest rate, both vowels and consonants were compressed by about 33% [24.38]. This shows that time-scale modifications should take into account phonetic information. Speaking at faster or slower rate again introduces modifications in pitch values since there are fluctuations in the subglottal pressure. This means that time-scale modifications should be performed jointly with pitch modifications.

In the source–filter theory presented at the beginning of the chapter, it was assumed that glottal airflow source is not influenced by the vocal tract. In reality, there is a nonlinear coupling between the source and the filter. Results from studies on the fine structure of the glottal airflow derivative waveform show that an increase in the first-formant bandwidth and modulation of the first-formant frequency occurs during the glottal open phase [24.40]. Obviously, when pitch modification is applied, these interactions should be respected.

Attempts have been made to incorporate some of these observations into the modification algorithms. In [24.41], a higher intelligibility score was achieved for time-scale-modified speech signals when nonstationarity measurements in the signal were taken into account. In [24.42], the interaction between pitch and spectral envelopes was modeled in a statistical way. This was used to postprocess pitch-modified signals. Perceptual tests have shown that this postprocessing improved the naturalness of the pitch-modified signal.

To improve further the quality of voice transformations more effort should be made taking into account nonlinear phenomena during the production process and results from the natural language processing area. In other words, voice transformation requires more than just modeling of the speech signal; it requires *understanding* of the speech process (production, perception, and language).

24.9 Summary

In this chapter, we have described voice transformations through a simple harmonic representation of speech. We began with the description of the basic source–filter theory for the production of speech and providing a mathematical description of speech production using this theory in the context of a harmonic model. We used this description to define voice transformation by specifying modifications for the source, for the filter, and their combination. We then provided formal definitions of these modifications and their application in the context of the harmonic model was also derived and a set of conditions for the pitch synchronous analysis of speech was described. Techniques for filter modifications were discussed and a state-of-the-art method based on a GMM description of the acoustic space of a speaker was developed. A mapping function that made use of the complete description of each component of the GMM was provided. Finally, we discussed speech quality issues related to voice transformations and noted that, for improving speech quality in the future, we need to give more realism to the source–filter model by taking into account the nonlinear coupling between the source and filter and to processes related to our perception system. In other words, just modeling the speech *signal* may be enough for transmission through the networks, but it is not sufficient for modifying the signal in a way that is perceived by humans to be natural. For this, we need to *understand* speech and then *develop* algorithms able to incorporate this understanding.

References

24.1 T. Dutoit: *An Introduction to Text-to-Speech Synthesis* (Kluwer Academic, Dordrecht 1997)

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

24.3 B. Atal, J. Remde: A new model of LPC excitation for producing natural-sounding speech at low bit rates, Proc. IEEE ICASSP, Vol. 7 (1982) pp. 614–617

24.4 M.R. Schroeder, B.S. Atal: Code-excited linear prediction (CELP): High-quality speech at very low bit rates, Proc. IEEE ICASSP, Vol. 10 (1985) pp. 937–940

24.5 R.J. McAulay, T.F. Quatieri: Speech analysis/synthesis based on a sinusoidal representation, IEEE ICASSP **34**, 744–754 (1986)

24.6 Y. Stylianou: Modeling speech based on harmonic plus noise models. In: *Nonlinear Speech Modeling and Applications*, ed. by G. Chellot, A. Esposito, M. Faundez (Springer, Berlin, Heidelberg 2005) pp. 375–383

24.7 A.V. Oppenheim, R.W. Schafer: Homomorphic analysis of speech, IEEE Trans. Audio Electroacoust. **16**, 221–228 (1968)

24.8 D.B. Paul: The spectral envelope estimation vocoder, IEEE ICASSP **29**, 786–794 (1981)

24.9 O. Cappé, E. Moulines: Regularization techniques for discrete cepstrum estimation, IEEE Signal Process. Lett. **3**(4), 100–102 (1996)

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

24.11 Y. Stylianou, J. Laroche, E. Moulines: High-quality speech modification based on a harmonic + noise model, Proc. Eurospeech, Vol. 95 (1995) pp. 451–454

24.12 T.F. Quatieri, R.J. McAulay: Shape invariant time-scale and pitch modification of speech, IEEE ICASSP **40**, 497–510 (1992)

24.13 R.J. McAulay, T.F. Quatieri: Low-rate speech coding based on the sinusoidal model. In: *Advances in Speech Signal Processing*, ed. by S. Furui, M. Sondhi (Marcel Dekker, New York 1991) pp. 165–208, Chap. 6

24.14 L. Almeida, F. Silva: Variable-frequency synthesis: An improved harmonic coding scheme., Proc. IEEE ICASSP, Vol. 9 (1984) pp. 437–440

24.15 E. Moulines, J. Laroche: Techniques for pitch-scale and time-scale transformation of speech. Part I. Non parametric methods, Speech Commun. **16**, 175–205 (1995)

24.16 E. Moulines, F. Charpentier: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Commun. **9**, 453–467 (1990)

24.17 S. Roucos, A. Wilgus: High-quality time-scale modification of speech, Proc. IEEE ICASSP (1985) pp. 493–496

24.18 W. Verhelst, M. Roelands: An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech, Proc. IEEE ICASSP, Vol. 2 (1993) pp. 554–557

24.19 M. Abe, S. Nakamura, K. Shikano, H. Kuwabara: Voice conversion through vector quantization, Proc. IEEE ICASSP, Vol. 1 (1988) pp. 655–658

24.20 K. Shikano, K. Lee, R. Reddy: Speaker adaptation through vector quantization, Proc. IEEE ICASSP, Vol. 11 (1986) pp. 2643–2646

24.21 H. Kuwabara, Y. Sagisaka: Acoustic characteristics of speaker individuality: Control and conversion, Speech Commun. **16**(2), 165–173 (1995)

24.22 N. Iwahashi, Y. Sagisaka: Speech spectrum transformation based on speaker interpolation, Proc. IEEE ICASSP, Vol. 1 (1994) pp. 461–464

24.23 H. Valbret, E. Moulines, J. Tubach: Voice transformation using PSOLA techinques, Speech Commun. **11**(2-3), 175–187 (1992)

24.24 H. Mizuno, M. Abe: Voice conversion algorithm based on piecewise linear conversion rule of formant frequency and spectrum tilt, Speech Commun. **16**, 153–164 (1995)

24.25 Y. Stylianou, O. Cappé, E. Moulines: Statistical methods for voice quality transformation, Proc. Eurospeech, Vol. 95 (1995) pp. 447–450

24.26 Y. Stylianou, O. Cappé, E. Moulines: Continuous probabilistic transform for voice conversion, IEEE Trans. Speech Audio Process. **6**(2), 131–142 (1998)

24.27 A. Kain, M. Macon: Spectral voice conversion for text-to-speech synthesis, Proc. IEEE ICASSP, Vol. 5 (1998) pp. 285–288

24.28 A. Mouchtaris, J.V. derSpiegel, P. Mueller: Non parallel training for voice conversion based on a parameter adaptation, IEEE Trans. Audio Speech Language Process. **14**(3), 952–963 (2006)

24.29 D. Suendermann, H. Hoege, A. Bonafonte, H. Ney, A. Black, S. Narayanan: Text-independent voice conversion based on unit selection, Proc. IEEE ICASSP, Vol. 1 (2006) pp. 81–84

24.30 R.O. Duda, P.E. Hart: *Pattern Classification and Scene Analysis* (Wiley, New York 1973)

24.31 R.C. Rose, D.A. Reynolds: Text independent speaker identification using automatic acoustic segmentation, Proc. IEEE ICASSP, Vol. 1 (1990) pp. 293–296

24.32 A.P. Dempster, N.M. Laird, D.B. Rubin: Maximum likelihood from incomplete data via the EM algorithm (methodological), J. R. Stat. Soc. B **39**(1), 1–22 (1977)

24.33 A.P. Dempster, N.M. Laird, D.B. Rubin: Maximum likelihood from incomplete data via the EM algorithm (discussion), J. R. Stat. Soc. B **39**(1), 22–38 (1977)

24.34 L.R. Rabiner, B.-H. Juang: *Fundamentals of Speech Recognition* (Prentice Hall, Upper Saddle River 1993)

24.35 S.M. Kay: *Fundamentals of Statistical Signal Processing: Estimation Theory*, PH Signal Process. Ser. (Prentice Hall, Upper Saddle River 1993)

24.36 C. Chatfield, A.J. Collins: *Introduction to Multivariate Analysis* (Chapman Hall, Boca Raton 1980)

24.37 Y. Stylianou: *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. Thesis (Ecole Nationale Supèrieure des Télécommunications, Paris 1996)

24.38 T.F. Quatieri: *Discrete-Time Speech Signal Processing* (Prentice Hall, Englewood Cliffs 2002)

24.39 J.M. Pickett: *The Sounds of Speech Communication* (Pro-Ed, Austin 1980)

24.40 C. Jankowski: *Fine Structure Features for Speaker Identification*, Ph.D. Thesis (Massachusetts Institute of Technology, Cambridge 1996)

24.41 D. Kapilow, Y. Stylianou, J. Schroeter: Detection of non-stationarity in speech signals and its application to time-scaling, Proc. Eurospeech, Vol. 99 (1999) pp. 2307–2310

24.42 A. Kain, Y. Stylianou: Stochastic modeling of spectral adjustment for high quality pitch modification, Proc. IEEE ICASSP, Vol. 2 (2000) pp. 949–952