Spectral and temporal modifications

Translation in English of Bertrand David's course handout

roland.badeau@telecom-paris.fr

Voir Page 18

Contents

This course handout draws from passages of various documents and mainly from a work specifically dedicated to audio signal processing [9] (Chap. 7). It develops more particularly the phase vocoder-based methods and the temporal methods.

Translation in English of Bertrand David's course handout
roland.badeau@telecom-paris.fr

Voir Page 18

Chapter 1

Spectral and temporal modifications

1.1 Introduction

The objective of these modifications, which correspond to usual needs in various fields of sound and speech processing, is to *independently* control the temporal, spectral and possibly formantic (slow variations of the spectrum) evolutions of the signal:

- temporal dilation: we want to modify duration scales without altering the spectral content and especially the pitch in the case of a harmonic signal,
- pitch variation: in the case of harmonic signals, we want to change the pitch of the sound, while retaining its temporal evolution (for example the prosodic flow in the case of speech) and especially its duration,
- formantic control: in the case of a pitch modification, one can choose to either modify the spectral scale as a whole (and therefore to move the formants or spectral envelope) or to keep the spectral envelope constant while transposing the line spectrum.

Applications where these types of independent modifications are numerous:

- synthesis by sampling of a wave table (musical sounds or speech segments [1]),
- post-synchronization: to perform a synchronization of sound and image,
- data compression [13]
- reading for blind people: our inner reading is much faster than our diction. By shrinking durations we can allow blind people to increase their speed of browsing documents,
- learning foreign languages: slowing the speech flow is helpful,
- musical post-production: to mix several recordings it may be useful to speed up or slightly reduce the tempo. It may also be interesting to locally correct the precision of a voice or instrument.

We can classify in three types the methods that carry out these modifications:

- methods inspired by the circular reader head or modified radiocassette (we add / subtract portions of the signal [6]). These methods are called *temporal*,
- phase vocoder-based methods (*spectral* methods using STFT [18, 20]),
- methods based on signal models (LPC [12], Sinus+Noise [23], Audio grains [8],...).

Temporal methods have resulted in many developments in digital signal processing: SOLA methods (Synchronized Overlapp Add) [19], PSOLA (Pitch Synchronized Overlapp Add) [16]. This last one uses a synchronized copy/deletion technique on the glottal impulses. This achieves a very good quality modification of the time scale *without resampling the signal*.

The PSOLA method can be adapted to perform formantic modifications by modifying the duration of the segments without modifying the position of the glottal impulses. In this way, we can transpose the spectral envelope without modifying the pitch or the duration, and thus modify the timbre of a voice (transform a male voice into a more female voice by example). Other techniques of formantic modifications use cepstral representations [3].

1.2 Signal models to define temporal and spectral distortions

A simple replay at 16 kHz of an audio signal sampled at 8 kHz is enough to convince us that the temporal and spectral expansions or compressions are interdependent. This dependence can be interpreted as a simple theoretical result on the Fourier transform (FT): the Heisenberg uncertainty relation translates this dependence in terms of supports, and the high frequency decrease in the FT is related to the regularity of the time signal. Therefore the definition of *independent* temporal and spectral distortions can only be obtained for well-defined signal *models*.

1.2.1 McAuley-Quatieri model

Speech production model. The most common and widely used speech signal model is that of a time-varying linear filter, excited by a harmonic source (in the case of voiced sounds) or by a stationary random process with flat spectrum (in the case of unvoiced sounds). In the present case, we consider the voiced case, for which this source is a sum of sinusoidal components whose frequencies are multiple of a fundamental frequency $f_0(t)$. This representation is equivalent to writing the source as a Dirac comb whose period depends on time.

Let $g_t(\tau)$ be the impulse response of the system at time t. The signal is then simply written as a function of the excitation signal $e(t)$:

$$x(t) = \int_{-\infty}^{+\infty} g_t(\tau) e(t-\tau) d\tau. \tag{1.1}$$

This non time-invariant system can be represented by a time-dependent frequency response:

$$G(t, f) = M(t, f) \exp j\varphi(t, f).$$

The temporal variations of g_t are linked to the articulatory movements and are considered slow compared to the fundamental period of the signal. On the other hand, these variations are assumed to be weak over the duration of the filter memory. The system is *quasi-stationary*.

For voiced speech, that is to say involving a periodic vibration of the vocal cords, the excitation signal writes:

$$e(t) = \sum_{k=-\infty}^{+\infty} \exp j\xi_k(t) \tag{1.2}$$

with

$$\xi_k'(t) = 2\pi f_k(t).$$

The quasi-stationary nature of g_t leads to a practical limitation of the support of this function to a dimension of the order of the system memory. The integral in expression 1.1 is therefore well-defined in practice. In the same way, the frequency support of speech is limited in practice and the discrete sum of expression 1.2 is in fact a finite sum of $L(t)$ complex exponential terms. Taking into account the fact that f_0 varies little over the memory duration of the filter we can expand

$$\xi_k(t-\tau) \approx \xi_k(t) - 2\pi\tau k f_0(t)$$

in the vicinity of t (*i.e.* for τ lower than the memory of the filter). Then we obtain:

$$x(t) = \sum_{k=1}^{L(t)} M(t, f_k(t)) \exp j[\xi_k(t) + \varphi(t, f_k(t))] \tag{1.3}$$

Mc-Auley and Quatieri model. This model was introduced by McAulay and Quatieri around 1985 [15], mainly for low rate speech coding. So it is related to the expression obtained in 1.3. It is however a little more general since it does not assume a necessarily harmonic relationship between the instantaneous frequencies. The signal is represented as a sum of sines whose frequencies, amplitudes and phases are controlled over time:

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t) \qquad \text{with} \qquad \Psi'_k(t) = \omega_k(t) = 2\pi f_k(t) \tag{1.4}$$

where $A_k(t)$ is the amplitude at time t of sine k, $\Psi_k(t)$ is the *instantaneous phase* of this sine at time t and $f_k(t)$ is its *instantaneous frequency*. This decomposition is not unequivocal and we generally consider that the functions $A_k(t)$ and $\omega_k(t)$ have slow variations compared to the functions $\exp(j\Psi_k(t))$.

1.2.2 Serra-Smith model

This model was developed in the early 90s [23] in order to meet the need for an analysis/synthesis system accounting for the noisy component of music or speech. This component is very expensive to represent as a sum of sines. The proposed model is therefore an extension of that of MacAuley-Quatieri:

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t) + b(t) \tag{1.5}$$

where $b(t)$ is a stationary random process filtered by a time-varying filter, like filter g_t presented above. Let h_t be this filter, we will then write, taking into account the causality of signals,

$$b(t) = \int_0^t h_t(\tau) u(t-\tau) d\tau \tag{1.6}$$

where $u(t)$ is a white stationary random process.

The complete analysis/modification/synthesis system includes

- an estimation phase of the deterministic components,
- a phase of linear interpolation of the amplitudes and cubic interpolation of the phases from one frame to another of the signal for these components,
- a subtraction of this deterministic part to get $b(t)$ for each frame,
- the application of a possibly distinct transformation algorithm for each of the two components,
- resynthesis.

1.3 Definitions and equivalences

All the definitions given here relate to a model of signal with sinusoidal components. They therefore apply to the McAuley-Quatieri model or to the deterministic part of the Serra-Smith model. The phases at $t = 0$ will be assumed to be zero for the sake of simplification (this term can be incorporated into the definition of the amplitudes).

1.3.1 Temporal distortion

We define the time distortion function using the new time scale τ and the original time scale t by:

$$\tau = T(t). \tag{1.7}$$

This function is continuous and bijective from $\mathbb{R}^+$ to $\mathbb{R}^+$. The modification of the signal's time scale $x(t)$ is then defined by

$$y(\tau) = \sum_{k=1}^{L(T^{-1}(\tau))} A_k(T^{-1}(\tau)) \exp(j\phi_k(\tau)) \tag{1.8}$$

The conservation of the frequency content then requires to maintain the values of the instantaneous frequencies, hence the relation:

$$\phi_k(\tau) = \int_0^{\tau} \omega_k(T^{-1}(u))du \tag{1.9}$$

1.3.2 Pitch modification

To modify the pitch of the signal $x(t)$ we build the signal:

$$y(t) = \sum_{k=1}^{L(t)} A_k(t) \exp(j\Phi_k(t)) \tag{1.10}$$

The alteration of the frequency content is defined using a function $\alpha(t)$ called frequency compression rate, according to the expression:

$$\Phi_k(t) = \int_0^{t} \alpha(u)\omega_k(u)du \tag{1.11}$$

1.3.3 Reciprocity

By a quick calculation we show that the operating sequence: $x \rightarrow x_1$ by time distortion ($\tau = T(t)$) followed a simple replay at a different temporal speed (*i.e.* without maintaining the frequency characteristics) $x_1(\tau) = y(v)$ with $v = T^{-1}(\tau)$ is equivalent to a frequency modification governed by the function $\alpha(t) = T'(t)$, that is to say:

$$y(v) = \sum_{k=1}^{L(v)} A_k(v) \exp j\Phi_k(v) \tag{1.12}$$

with

$$\Phi_k(t) = \int_0^{v} T'(u)\omega_k(u)du \tag{1.13}$$

This relationship is particularly useful in cases where the corresponding time distortion is a multiplying factor, like for instance $T(t) = 2t$. Then $T'(t)$ is constant and the replay operation is a simple replay of the signal obtained at a different rate (for example, for sampled signals, $F'_e = 2Fe$ in the previous case).

1.4 Short-term Fourier transform

The methods of analysis/synthesis and modification of sounds based on the use of the Short Term Fourier Transform (STFT) are very common. The corresponding tool is usually called *phase vocoder*. It refers to the polar representation (module & phase) of the STFT.

1.4.1 Theoretical reminders

The block diagram of the STFT is represented in figure 1.1, as it is numerically computed. The principle is that of a sliding Fourier transform, performed on overlapping frames of the signal. Each frame is windowed by an analysis window. We will write the STFT of a digital signal $x(n)$ in the form

$$\tilde{X}(t_a, \nu) \stackrel{\Delta}{=} \sum_{n\in\mathbb{Z}} x(n + t_a)w_a(n)e^{-j2\pi\nu n}. \tag{1.14}$$

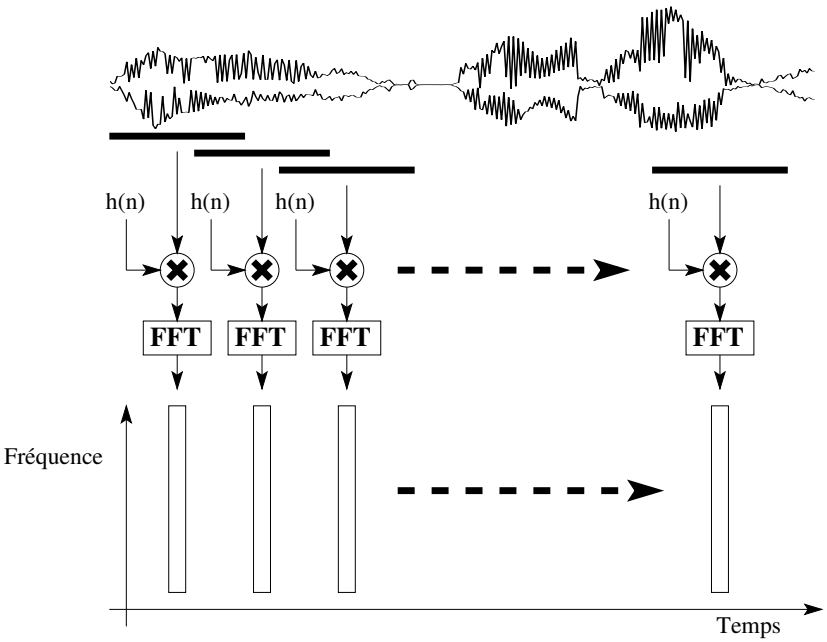

Figure 1.1: Short-term Fourier transform

w_a denotes the analysis window, most often of finished length, real and symmetrical. The analysis times are implicitly indexed by a natural integer u, that is $t_a = t_a(u), \quad u \in \mathbb{N}$. We preferred here a notation function of ν while a notation function of $e^{j2\pi\nu}$ would have been more consistent with the interpretation in terms of sliding Fourier transform, but this choice simplifies the expressions.

Interpretation. A quick calculation shows that, by defining $h(n) = w_a(-n)e^{j2\pi\nu_p n}$, the expression 1.14 can be written in the form of a convolution product:

$$\tilde{X}(t_a, \nu_p) = [x * h](t_a). \tag{1.15}$$

If $w_a(n)$ is a real and pair window of finite length, its FT $W_a(e^{j2\pi\nu})$ is real and even. The FT of h is then simply $H(e^{j2\pi\nu}) = W_a(e^{j2\pi(\nu-\nu_p)})$. An example of typical result is given in figure 1.2 for $\nu_p = 0.3$. This example shows that $\tilde{X}(t_a, \nu_p)$ performs a bandpass FIR filtering around frequency ν_p. The characteristics of the filter are linked to that of the chosen analysis window. This interpretation is at the origin of the qualification of *bandpass convention* given to expression 1.14. There is another convention, called *low pass*, often used for its ease of handling calculations. We will, however, stick to the band pass convention because it corresponds to the practical realization.

Discrete version of the STFT. In practice, the Fourier transform is evaluated using the DFT. This is equivalent to setting $\nu_p = p/N$ in the expression of $\tilde{X}(t_a, \nu_p)$. N is the order of the DFT. We thus obtain a discrete version of the STFT, i.e. sampled in frequency, i.e.

$$\tilde{X}(t_a, \nu_p) = \sum_{n=0}^{N-1} x(n + t_a)w_a(n)e^{-j2\pi\frac{pn}{N}}. \tag{1.16}$$

In order to avoid time aliasing, the length of the analysis windows will be less than or equal to N.

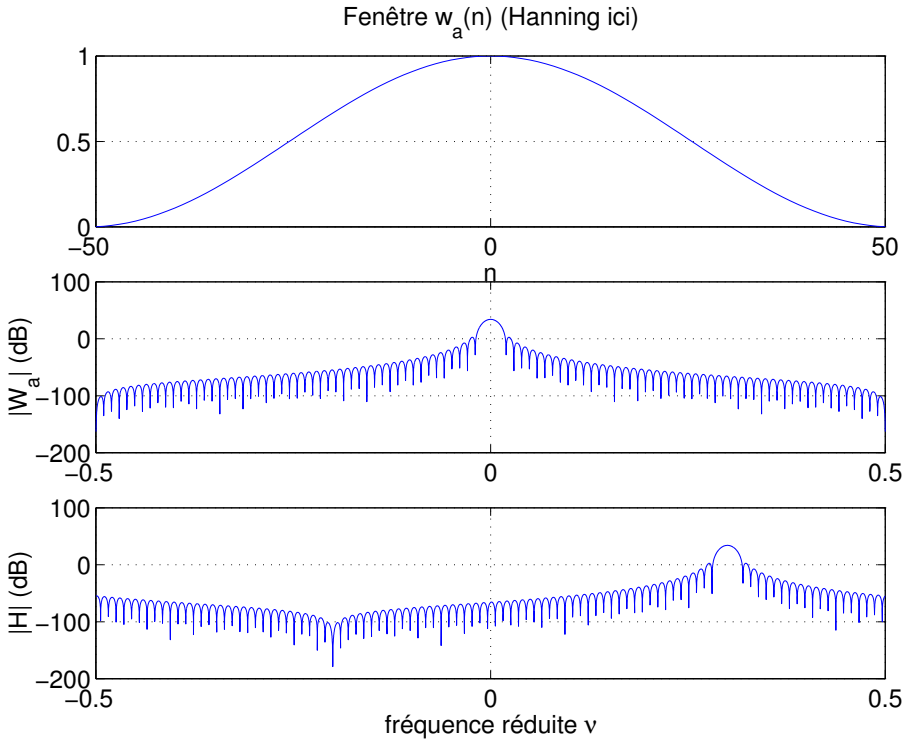

Figure 1.2: Bandpass filtering equivalent to an STFT channel

Modifications and problems posed. The modification of sounds by the phase vocoder involve obtaining a modified STFT from $\tilde{X}(t_a, \nu_p), \quad k = 0, \ldots, N-1$, then resynthesizing the signal. We denote by $t_s = t_s(u)$ the temporal synthesis marks. Hence the modification

$$\tilde{X}(t_a(u), \nu_p) \rightarrow Y(t_s(u), \nu_p).$$

The main difficulty encountered is that Y must satisfy strong conditions [18] to correspond to a well-defined original sequence. The solution to this problem is found in the least squares sense [17]. However, one can write the perfect reconstruction conditions in case no modification is performed (*i.e.* $t_s = t_a$ and $Y = \tilde{X}$).

Perfect reconstruction condition. The reverse operation of the analysis is carried out for the synthesis: from the stream of discrete spectra $\tilde{X}(t_a(u), \nu_p)$ we compute an inverse DFT and we reconstruct the signal by overlapp-add (or OLA). The result is given by

$$y(n) = \sum_u w_s(n - t_s(u)) y_w(n - t_s(u), t_s(u)) \tag{1.17}$$

with $t_s(u) = t_a(u)$ and

$$y_w(n, t_s(u)) = \frac{1}{N} \sum_{p=0}^{N-1} Y(t_s(u), \nu_p) e^{j2\pi\nu_p n}.$$

Taking $Y = \tilde{X}$ into account and substituting the expression 1.14 in 1.17, we show that $x(n) = y(n)$ is obtained by using the sufficient condition:

$$\sum_u w_a(n - t_a(u)) w_s(n - t_a(u)) = 1 \tag{1.18}$$

1.5 Modifications using the phase-vocoder

1.5.1 Instantaneous frequency

The transformations presented in section 1.3 require the calculation of the instantaneous frequencies $\omega_k(t)$ of each of the components of the sum 1.4. This calculation is carried out from two successive short term spectra $\tilde{X}(t_a(u), \nu_p)$ and $\tilde{X}(t_a(u+1), \nu_p)$ $p = 0, \ldots, N-1$, under certain conditions that ensure the existence of a solution.

Narrow-band condition. This first condition ensures the presence of *at most* one component per channel of the STFT. The substitution of the expression 1.4,

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t),$$

into expression 1.14 of the STFT gives:

$$\tilde{X}(t_a(u), \nu_p) = \sum_{n=0}^{N-1} \sum_{k=1}^{L(n+t_a)} A_k(n+t_a) \exp(j\Psi_k(n+t_a)) w_a(n) e^{-j2\pi\nu_p n}$$

We then use the quasi-stationarity of the model, namely:

$$\begin{aligned} A_k(n+t_a) &\approx A_k(t_a) \\ \Psi_k(n+t_a) &\approx \Psi_k(t_a) + n\omega_k(t_a) \end{aligned}$$

and finally, by defining $\omega_k(t) = 2\pi f_k(t)$:

$$\tilde{X}(t_a(u), \nu_p) = \sum_{k=1}^{L(n+t_a)} A_k(t_a) \exp(j\Psi_k(t_a)) W_a(e^{j2\pi(\nu_p - f_k(t_a))}) \tag{1.19}$$

The narrow band condition leads to non-negligible values of $W_a(e^{j2\pi(\nu_p - f_k(t_a))})$ for at most one value of k. Let $k = l$ be this value if it exists. If we note f_c the cutoff frequency of the low-pass filter whose impulse response is $w_a(n)$, then the existence of l implies

$$|\nu_p - f_l(t_a)| \leq f_c \ ,$$

that is, the component number l is in the pass-band of the filter corresponding to the p-th channel of the STFT. The expression 1.19 is then reduced to the contribution of the l-th component alone:

$$\tilde{X}(t_a(u), \nu_p) = A_l(t_a) \exp(j\Psi_l(t_a)) W_a(e^{j2\pi(\nu_p - f_l(t_a))}) \tag{1.20}$$

If we assume that w_a is real and even, and therefore W_a is real and even, this expression is interpreted as follows: the phase of the STFT gives access to the instantaneous phases of the components of $x(t)$, up to an indeterminacy of a multiple of 2π, and the module of the STFT gives access to instantaneous amplitudes of $x(t)$, up to an amplitude factor due to filtering. We can therefore deduce the instantaneous frequencies of each component from the phase of the flow of short-term spectra, provided that the indeterminacy of 2π is removed.

Example: for a Hann analysis window of length L, the narrow-band condition applied to a line spectrum of harmonics (case of a voiced speech segment for example) leads to a spacing of spectral peaks at least equal to the bandwidth of the Fourier transform of the window, that is $4/L$. That results in $f_0 < 4/L$, i.e. a window length at least equal to 4 times the fundamental period.

Overlap condition. We will see here that the removing of the indeterminacy leads to a condition of minimal recovery of the analysis windows. Indeed, the phase difference between two successive analysis times, for the p-th channel of STFT is written, by defining $\Phi(t_a(u), \nu_p) = \arg \tilde{X}(t_a(u), \nu_p)$,

$$\begin{aligned} \Delta\Phi_p &= \Phi(t_a(u+1), \nu_p) - \Phi(t_a(u), \nu_p) = \Psi(t_a(u+1)) - \Psi(t_a(u)) \ [2\pi] \\ &= 2\pi f_l \Delta t_a(u) + 2n\pi \\ &= 2\pi(f_l - \nu_p)\Delta t_a(u) + 2\pi\nu_p \Delta t_a(u) + 2n\pi \end{aligned}$$

Translation in English of Bertrand David's course handout
roland.badeau@telecom-paris.fr

where n is a relative integer and $\Delta t_a(u) = t_a(u+1) - t_a(u)$. By taking $|\nu_p - f_l(t_a)| \leq f_c$ into account, the previous equation leads, if the condition 1.21 below is verified

$$f_c \Delta t_a(u) < 1/2, \tag{1.21}$$

to the inequality

$$|\Delta\Phi_p - 2\pi\nu_p \Delta t_a(u) - 2n\pi| < \pi \ ,$$

however there is one and only one value of n that verifies this property. This has removed the indeterminacy. In summary, we can thus get the value of the instantaneous frequency *in each STFT channel* by the following algorithm:

1. Calculation of the STFT at two successive times of analysis, which gives $\Delta\Phi_p$ for each channel ($p = 0, \ldots, N-1$),
2. For each channel, we look for the value $Q(n_0)$ of $Q(n) = \Delta\Phi_p - 2\pi\nu_p\Delta t_a - 2n\pi$ such that $|Q(n_0)| < \pi$,
3. we deduce the instantaneous frequencies by $f_l = \nu_p + \dfrac{Q(n_0)}{2\pi\Delta t_a}$.

Interpretation: inequality 1.21 leads to a minimum overlap condition between the analysis windows. Indeed, if we take for example a Hann window, for which we can estimate $f_c = 2/L$ where L is the window length, it becomes

$$\Delta t_a < \frac{L}{4}$$

which corresponds to a minimum recovery of 75% in analysis.

1.5.2 Temporal distortion

Once the instantaneous frequencies in each channel are deduced[1], the temporal signal distortion may be considered. In particular, instantaneous phases can be "unwound" so as to synchronize the modified STFT on the synthesis times. We then obtain the following modification algorithm, assuming that the analysis STFT $\tilde{X}(t_a(u), \nu_p)$ and the synthesis STFT $\tilde{Y}(t_s(u), \nu_p)$ are calculated for the index u, and given the time distortion law $T(t)$:

1. calculation of the STFT at time $t_a(u+1)$ and deduction of the instantaneous frequency $f_k(t_a(u))$ in each channel,
2. calculation of the new synthesis time $t_s(u) = T(t_a(u))$; in practice we take the whole part of this new instant
3. iteration of the synthesis instantaneous phase

$$\Phi_s(t_s(u+1), \nu_p) = \Phi_s(t_s(u), \nu_p) + 2\pi f_p(t_a(u))(t_s(u+1) - t_s(u))$$

4. calculation of the synthesis STFT for the index $u+1$ according to

$$\tilde{Y}(t_s(u+1), \nu_p) = A_p(t_a(u+1)) \exp j\Phi_s(t_s(u+1), \nu_p)$$

1.5.3 Pitch modification

The modification of pitch, or more generally of the frequency scale, is obtained either by temporal resampling or by spectral resampling.

[1] in doing so, we assume that there is one and only one component by channel and therefore we can identify the indexes p (STFT channel) and k (components).

Temporal resampling. This method is based on the reciprocity properties as seen in paragraph 1.3.3.

In the case of a constant frequency compression ratio $\alpha(t) = \alpha_0$, we obtain the desired modification by

1. a time stretch of factor α_0,
2. a replay at sampling frequency $\alpha_0 F_e$

where F_e is the original sampling frequency. This technique is equivalent to performing a resampling of factor $\alpha_0 = F'_e/Fe$ and playing at F_e. In this last case, it should however be noted that the time support is divided by α_0.

An extension of this resampling technique can be applied to obtain compression ratios $\alpha(t)$ variable over time. We use the canonical resampling method of digital signals by approaching α by a rational fraction at each analysis time $\alpha(t_a) = L(t_a)/M(t_a)$ and by performing the processing chain of figure 1.3 where $H(z)$ is a low-pass filter of cutoff frequency $\nu_c = \min(1/2L, 1/2M)$. We can therefore apply this processing to each frame of the signal and

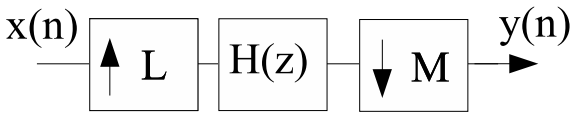

Figure 1.3: Canonical resampling chain of factor L/M

use synchronized analysis and synthesis times $t_a(u) = t_s(u)$. It should be noted that in this case, the phase vocoder is not used. This method can be quite demanding because it requires the calculation of a new interpolation filter H at each analysis step.

Spectral resampling. The phase vocoder allows a less greedy solution than the resampling in the case of variable compression ratios, by performing resampling in the frequency domain. This frequency resampling is performed by linear interpolation of the analysis short-term spectrum, that is

$$\begin{aligned} q &= \lfloor p/\alpha(t_a(u)) \rfloor \\ \mu_p &= p/\alpha(t_a(u)) - q \\ \tilde{Y}(t_s(u), \nu_q) &= (1-\mu_p)\tilde{X}(t_a(u), \nu_p) + \mu_p \tilde{X}(t_a(u), \nu_{p+1}) \end{aligned}$$

p and q are natural integers that index the channels of the STFT, they therefore vary from 0 to $N-1$. We note that this interpolation, if it presents no difficulty for compression rates greater than unity (pitch is increased), however requires completion of the high frequency spectrum for rates lower than unity (the synthesized sound is lower-pitched). One way to achieve this completion was suggested in [22] and simply consists of copying the low frequency part of the spectrum into the missing part. This spectral copy gives good rendering for sampling frequencies of at least 16 kHz. In this case, the completion occurs at high frequencies where the sound has mainly unvoiced characteristics.

Finally, to carry out the modification, it is necessary to take into account the local modification of the time scale caused by the frequency modification. Indeed the phases of the synthesis STFT $\Phi_s(t_s(u), \nu_p) = \Phi_a(t_a(u), \nu_p)$ are now synchronized on synthesis times different from the analysis times:

$$\begin{aligned} \Phi_s(t_s(u+1), \nu_p) &= \Phi_s(t_s(u), \nu_p) + 2\pi f_p(t_a(u))\Delta t_a(u) \\ &= \Phi_s(t_s(u), \nu_p) + 2\pi\alpha(t_a(u)) f_p(t_a(u))\Delta t_s(u) \end{aligned}$$

We therefore see that the local analysis duration $\Delta t_a(u)$ has been divided by α in the synthesis. Let $\Delta t_s(u) = \Delta t_a(u)/\alpha(t_a(u))$. This corresponds to a virtual time distortion

$$T(t) = \int_0^t \alpha(w)^{-1} dw$$

To carry out the time scale modification, it is therefore necessary to finally apply a compensatory temporal distortion $D(t) = T^{-1}(t)$.

Note regarding the processing of spoken or singing voice. In the case of pitch modifications in spoken or singing voice, a direct transposition of the signal leads to the "Donald Duck" effect. Indeed, the transposition of the overall spectrum leads to a transposition of its envelope and therefore of the formants. The timbre is then severely modified and the voice acquires a nasal characteristic evoking the sound of the duck. This effect is also produced by the modification of the characteristic impedance of the medium caused by the mixed gaz breathed in by divers. A solution to overcome this defect consists in estimating the spectrum envelope before the processing (by LPC or direct modeling [5]). The processing is then applied to the source signal (LPC residual for instance) then the obtained result is filtered to find the original spectral envelope, unchanged.

1.6 Pitch synchronous temporal method

This method (called TD-PSOLA, Time-Domain Pitch-Synchronous Overlap-Add) assumes that we process a speech signal whose period is known.

The idea [16] is based on the assumption that the speech signal is made up of glottal pulses filtered by the vocal tract. We thus observe a succession of impulse responses, positioned at multiple times of the period (assumption of the time comb convoluted with the impulse response of the vocal tract).

We then define "analysis marks" synchronous with the fundamental frequency for the voiced parts, positioned on the waveform at each period. Scale modifications are then carried out as follows:

1.6.1 Modification of the time scale.

In order to modify the signal duration without altering the fundamental frequency, we will simply duplicate (for time stretching) or eliminate (for time compression) periods of the waveform, depending on the desired modification rate. So we are led to define synthesis marks also synchronous with the fundamental period, associated with the analysis marks (in a non-bijective way since some marks are duplicated or eliminated).

Short-term signals around each analysis mark are then extracted (by the use of a time window, by example a Hann window, of duration equal to two periods and centered on the analysis mark) and 'copied' around the corresponding synthesis marks, and the modified signal is obtained by a simple OverLap-Add method. Figure 1.4 illustrates the principle of this method for a local time stretch rate of 1.5.

We see that two periods of the original signal gave birth to three periods in the modified signal, which corresponds well to a time stretch but the duration of the period is not not modified (the spacing of the synthesis marks is the same as that of the analysis marks), the fundamental frequency of the signal is preserved. Figure 1.6 gives an application example to the sentence "il s'est" whose original is given in figure 1.5. We notice the unvoiced part in the center of the window (the sound 's'), separating the two voiced parts /i/ and /e/.

1.6.2 Modification of the frequency scale.

If we are able to position the analysis marks in the signal exactly at the start of each glottal wave (impulse response of the vocal tract occurring at each glottal closure), we can see that decreasing (resp. increasing) the time interval separating two consecutive analysis marks will increase (resp. decrease) the fundamental frequency, without the formants being modified (the impulse response is not modified, in particular its temporal decay and its resonance frequencies - the formants).
We are thus led to define synthesis marks corresponding to the modified value of the fundamental, and to associate them with the analysis marks, as previously. Since the synthesis marks are closer (elevation of the fundamental) or farther (lowering of the fundamental) than in the original signal, we have to duplicate or eliminate some marks in order to keep the duration of the signal. Figure 1.7 illustrates the principle of this method.

It can be seen that the synthesis marks being more spaced out than the analysis marks, the signal period is lengthened. In order to avoid an elongation of the signal, it is necessary to periodically eliminate some short-term signals.

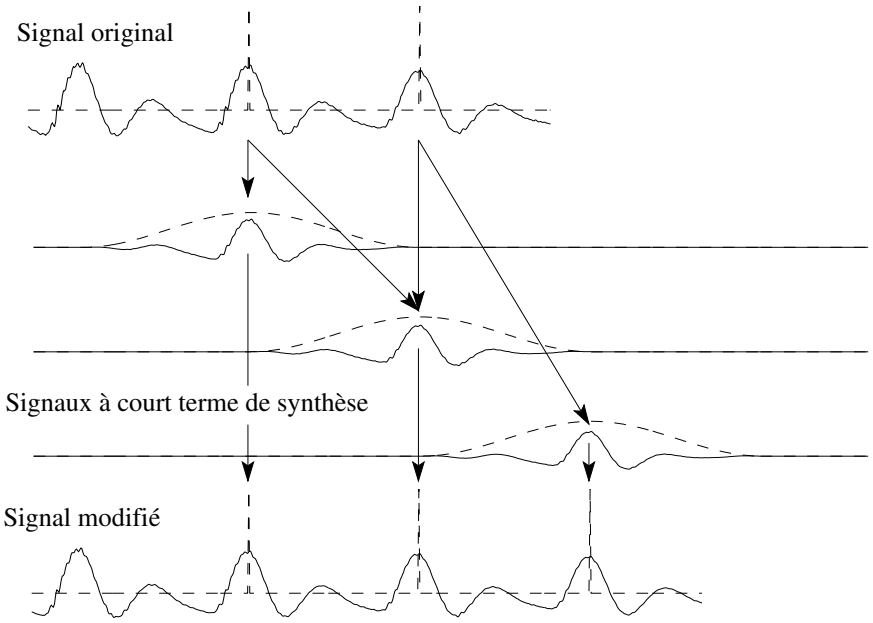

Figure 1.4: Modification of the duration of the signal by the TD-PSOLA method. At the top, the original signal, in the middle three short-term signals generated from two short-term signals centered around the first two analysis marks. At the bottom, the modified signal.

When the signal no longer has a precise fundamental frequency (case of consonants for instance), the modification is carried out non-synchronously, until we find a region with a sharper fundamental.

The method described above is mainly applied to speech, and makes very good quality modifications. By its simplicity, it can be subject to a real-time implementation. However, its application to more complex sounds, or sounds devoid of "pitch" (case of music in general) poses serious problems.
The modifications of fundamental frequency are however very sensitive to the position of the analysis marks. To make the method more robust, the modifications of frequency scale can be performed in the frequency domain (FD-PSOLA method) [16, 17].

For other methods based on very similar ideas, we can refer to [7, 14, 21, 26].

1.6.3 The circular memory technique

The circular memory technique is the simplest and most ancient time and frequency scale modification technique [2]. It is also a method operating in the time domain.

1.6.3.1 The analog origin

This technique is derived from an analog system proposed in the 1950s [6]. It consists in using a tape recorder equipped with a rotating head. The closed loop tape wraps around half of the cylinder (as for the VCR and the DAT) and scrolls at constant speed. The cylinder is provided with two diametrically opposed reader heads whose signals are mixed with an identical gain. It is possible to control the direction of rotation and the speed of the cylinder.

When the cylinder is motionless, the tape scrolls in an identical manner in front of the recording head and in front of one of the reader heads. The signal read is therefore identical to the signal recorded (up to recording

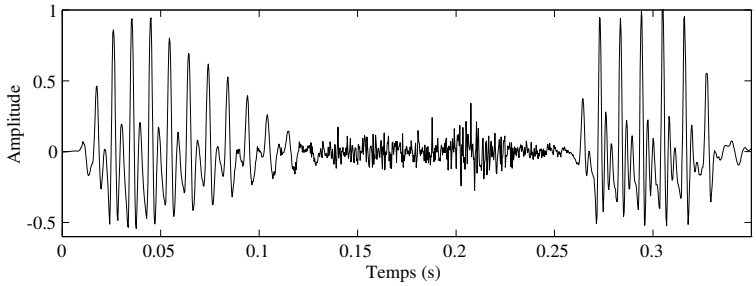

Figure 1.5: Original: "il s'est".

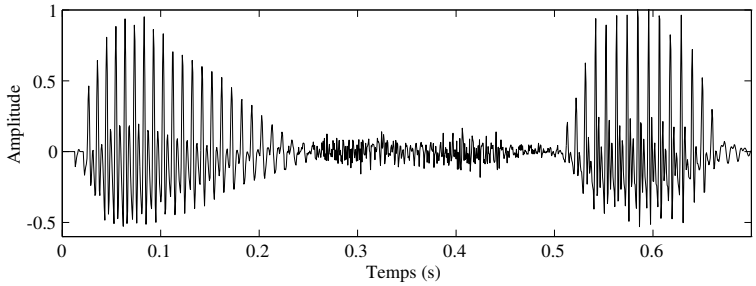

Figure 1.6: Signal stretched by factor 2.

errors).
When the cylinder rotates in the opposite direction of the scrolling of the tape, the relative speed V_r of the scrolling tape with respect to the reader head is faster than its absolute scroll speed V_a. During the period of contact between the reader head and the tape, the signal is thus read faster than it has been recorded, which corresponds to a dilation of the frequency axis. The presence of two heads ensures the continuity thanks to a natural cross-fade (when a head leaves the band, the other approaches it, so that the total signal does not decrease). Note that some portions of the signal can be read *two or more times*, depending on the speed of rotation of the head. It is this rereading that keeps the signal duration.
Conversely, when the cylinder rotates in the scrolling direction of the band, the frequency content of the signal is contracted towards the origin since the tape is read at a slower speed than it is recorded. In this case, some portions of the signal may not be read at all.
The ratio of frequency homothety is expressed as:

$$\alpha = \frac{V_r}{V_a} = \frac{V_a + R\,\Omega_{cylinder}}{V_a}$$

where V_a is the tape scrolling speed in front of the recording head, V_r is the relative speed of the tape with respect to the reader head, $\Omega_{cylinder}$ is the speed of rotation of the cylinder in radians s^{-1}, and R is the radius of the cylinder. In all cases, the regular alternation of the two heads induces a periodic "noise" of frequency $\Omega_{cylinder}/\pi$.

Modifications in the signal time scale are obtained for instance by recording the signal a first time on the tape, then by replaying it with a tape scrolling speed multiplied by factor α. In the absence of rotation of the reader head, the signal pitch is of course multiplied by factor α, which we try to avoid. We therefore compensate for the pitch change by a proper rotation of the reader head.

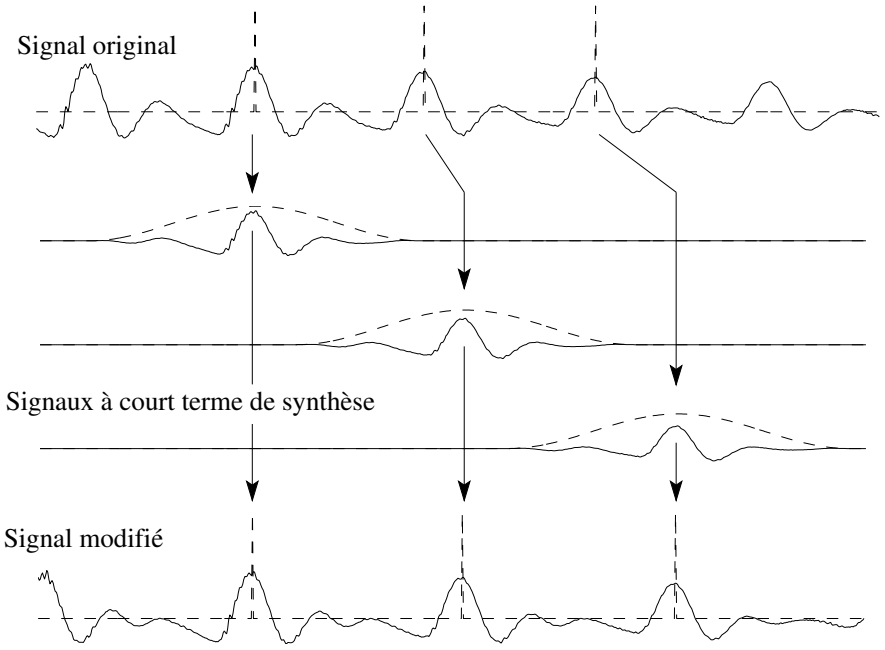

Figure 1.7: Modification of the pitch of the signal by the TD-PSOLA method. At the top, the original signal, in the middle three short-term signals generated from the three first analysis marks. At the bottom, the modified signal. The spacing of the synthesis marks is not identical to that of the analysis marks.

1.6.3.2 Digital implementation

Most commercially available pitch modifiers are based on a digital realization of the system described above. The magnetic tape is replaced by a circular memory in which the input signal samples are placed. This circular memory is read by two diametrically opposite pointers.

For each sample written in the memory (every ΔT seconds), we advance the reading pointers by $\alpha\Delta T$ seconds, where α is the rate of change, and then we read a sample in memory. In general (for non-integer values of α), we find ourselves between two samples, and as in the case of the "flanger", it is necessary to calculate the signal value at this time. Here too, a simple linear interpolation is suitable.
Thus, the signal is read with a sampling frequency different from the one it was recorded with, which causes a modification of the frequency scale of rate α. A problem arises when the reader pointer catches up (when $\alpha > 1$) or is caught (when $\alpha < 1$) by the writing pointer. As in the analog equivalent, continuity is ensured by a mixture of the two pointers at the time when the encounter occurs ("cross-fade"): the sample read by the current reader

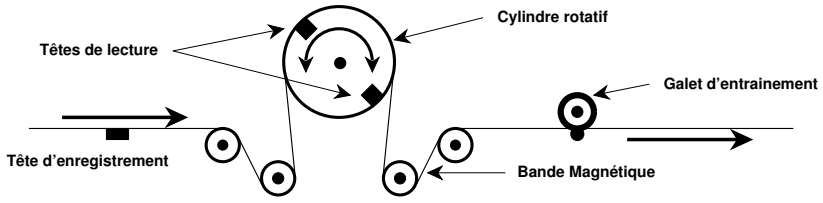

Figure 1.8: The circular memory technique

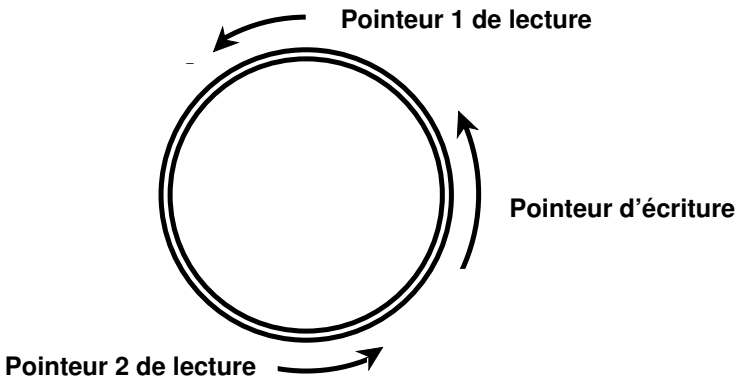

Figure 1.9: Digital implementation

pointer (e.g. pointer 1) is lowered while the one read by the other reader pointer (pointer 2) is increased. Finally, the second pointer becomes the current pointer, and keeps its maximum weighting, until the writing pointer gets close to it.

Implemented in this way, the pitch shifter has a behavior substantially equivalent to its analog counterpart (except that it is more easily configurable). Its implementation in real-time is not a particular problem, since it requires very few calculations.
Unfortunately, it produces artificial noise which comes from the periodic mixing of the two reader pointers. To try to improve the obtained quality, we seek to better combine the signals read by the two reader pointers, somewhat similarly to what is done in the synchronous methods. We can for example use the signal autocorrelation function to determine the most suitable location for the "cross-fading" [4, 10].

1.6.3.3 Modification of the duration by the technique of circular memory

Like its analog counterpart, the circular buffer technique can also be used for temporal scale modification: if you have a pitch shifter with circular memory, and want to perform a "time scaling" of parameter α, just change the signal sampling frequency by a rate α, then process it with the pitch shifter. So, in order to slow down the signal twice, it suffices to oversample it twice. If we listen to the signal obtained at the original frequency, it will be twice as much long, but also at a lower octave. So just listen to it at the original frequency by inserting a pitch shifter at rate $\alpha = 2$.

We quickly realize that it is easier to do both operations jointly: the technique then consists in repeating or periodically eliminating signal portions so as to increase (or decrease) the duration. Viewed from this angle, this technique (which is called "splicing method") approximates a TD-PSOLA technique in which we would not know the value of the fundamental frequency. The artifacts inherent in this method, which come from the breaks in the periodicity of the signal during the repetitions or eliminations, can be considerably reduced by the use of methods based on the autocorrelation of the signal in order to optimize the length and location of signal portions to be duplicated or destroyed [4, 10, 11, 19, 24, 25].

Bibliography

[1] J. Allen. Overwiew of text-to-speech systems. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, chapter 23, pages 741–790. Marcel Dekker, 1991.

[2] J. Benson. *Audio Engineering Handbook*. mcGraw-Hill, New York, 1988.

[3] O. Cappé, J. Laroche, and E. Moulines. Regularized estimation of cepstrum envelope from discrete frequency points. *IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, 1995.

[4] J. Dattorro. Using digital signal processor chips in a stereo audio time compressor/expander. *Proc. 83rd AES Convention, New York*, Oct 1987. preprint 2500 (M-6).

[5] A. El-Jaroudi and J. Makhoul. Discrete all pole modeling. *IEEE Trans. Acoust., Speech, Signal Processing*, 39(2):411–423, Fev 1991.

[6] G. Fairbanks, W.L. Everitt, and R.P. Jaeger. Method for time or frequency compression-expansion of speech. *IEEE Trans. Audio Electroacoust.*, AU-2:7–12, Jan 1954.

[7] E. Hardam. High quality time scale modification of speech signals using fast synchronized overlap add algorithms. *Proc. IEEE ICASSP-90*, pages 409–412, 1990.

[8] D.L. Jones and T.W. Parks. On the generation and combination of grains for music synthesis. *Computer Music J.*, 12(2):27–34, Summer 1988.

[9] M. Kahrs and K. Brandenbourg. *Applications of Digital Signal Processing to Audio and Acoustics*. Kluwer Academic Press, Dortrecht, Netherland, 1998.

[10] J. Laroche. Autocorrelation method for high quality time/pitch scaling. *IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, 1993.

[11] F. Lee. Time compression and expansion of speech by the sampling method. *J. Audio Eng. Soc.*, 20(9):738–742, 1972.

[12] J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63(11):1380–1418, Nov 1975.

[13] J. Makhoul and A. El-Jaroudi. Time scale modification in medium to low rate speech coding. *Proc. IEEE ICASSP-86*, pages 1705–1708, 1986.

[14] D. Malah. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Trans. Acoust., Speech, Signal Processing*, 27(2):121–133, 1979.

[15] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754, Aug 1986.

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

Translation in English of Bertrand David's course handout
`roland.badeau@telecom-paris.fr`

[17] E. Moulines and J. Laroche. Non parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16:175–205, Feb 1995.

[18] M. R. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24(3):243–248, Jun 1976.

[19] S. Roucos and A. M. Wilgus. High quality time-scale modification of speech. *Proc. IEEE ICASSP-85, Tampa*, pages 493–496, Apr 1985.

[20] M.R. Schroeder, J.L Flanagan, and E.A. Lundry. Bandwidth compression of speech by analytic-signal rooting. *Proc. IEEE*, 55:396–401, Mar 1967.

[21] R. Scott and S. Gerber. Pitch-synchronous time-compression of speech. *Proceedings of the Conference for Speech Communication Processing*, pages 63–65, Apr 1972.

[22] S. Seneff. System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24:358–365, 1982.

[23] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music J.*, 14(4):12–24, Winter 1990.

[24] B. Sylvestre and P. Kabal. Time-scale modification of speech using an incremental time-frequency approach with waveform structure compensation. *Proc. IEEE ICASSP-92*, pages 81–84, 1992.

[25] W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. *Proc. IEEE ICASSP-93, Minneapolis*, pages 554–557, Apr 1993.

[26] J.L. Wayman and D.L. Wilson. Some improvements on the sychronized-overlap-add method of time scale modification for use in real-time speech compression and noise filtering. *IEEE Trans. Acoust., Speech, Signal Processing*, 36(1):139–140, Jan 1988.