Waveform Preserving Time Stretching and Pitch Shifting for Sinusoidal Models of Sound

Riccardo Di Federico
Centro di Sonologia Computazionale, Dipartimento di Elettronica e Informatica
Università degli Studi di Padova
rdf@csc1.unipd.it http://www.dei.unipd.it/ricerca/csc/people/all/Di_Federico.html

Abstract

A method for performing waveform invariant time stretching and pitch shifting on a quasi harmonic and sinusoidally modeled sound is presented. The method is based on the *relative phase delay* representation of the phase, defined as the difference between the phase delay of the partials and the phase delay of the fundamental. This representation makes the waveform characterization independent from the phase of the first partial. It is therefore possible to compute a smooth trajectory for the phase of the modified fundamental and rebuild the waveform on the synthesis frame boundaries by adding the relative phase delays to the new fundamental phase delay.

1 Introduction

Traditional time stretching techniques which make use of time - frequency models often neglect phase, assuming that ear is sensitive only to frequency and amplitude of the partials. When the signal is not perfectly periodic, this assumption results in a waveform dispersion which gives the sound a 'phasy' or reverberant quality. In order to overcome this limit McAulay and Quatieri [2] [6] proposed to model the input signal by the sinusoidal version of the classical linear speech production model [8], in which the signal is interpreted as a pulse-like excitation passed through a linear system implementing the vocal tract frequency response. Time and pitch modifications are then obtained by modifying the excitation pulse onsets.

Here a different approach is proposed, based directly on waveform preservation rather than on source – filter separation, thus avoiding the spectral deconvolution process.

The basic idea is to force the waveform at the synthesis (scaled) frame boundaries to be the same as in the analysis frame. For this purpose a slightly modified representation of the sinusoidal parameters has been introduced. Besides amplitudes and frequencies, phases have been characterized by phase delays (phase / radian frequency ratios) instead of absolute values (radians). This allows to establish a relation among the time positions of the partials which is independent from the current time reference (the frame boundary time instant). When performing time stretching, partial amplitudes and frequencies are left untouched, while phase evaluation at time-scaled frame boundaries proceeds as follows. The phase of the first partial is updated from the previous synthesis frame by a propagation formula which adds the original phase variation scaled by the time stretching factor. The phases of the remaining partials are set so as to replicate the phase delay differences found in the corresponding analysis frame. Time scaled frames are then interpolated to generate a constant output frame rate.

The presented method proves to be simple, accurate and robust, while maintaining a high sound quality.

The paper is organized as follows. After a short description of the sinusoidal framework (Section 2), the *relative phase delay* model is introduced (Section 3) and used for the formulation of the time stretching method (Section 4). Finally, a straight extension of the algorithm to pitch shifting is introduced.

2 Sinusoidal model

The sinusoidal model [1] [3], assumes that a signal $s(t)$ can be approximated by a sum of N sinusoids with time varying parameters:

$$s(t) = \sum_{k=1}^{N} A_k(t)\cos[\theta_k(t)] + r(t) \qquad (1)$$

where the phase $\theta_k(t)$ of the k-th sinusoid is the integral of the time varying radian frequency $\omega_k(t)$:

$$\theta_k(t) = \int_0^t \omega_k(u)\, du + \theta_k(0) \qquad (2)$$

and where the residual $r(t)$, which should contain the low level noisy part of the sound, is supposed, for the purpose of this work, to be negligible (as in [1]). Processing of the residual is however an important issue [3] and will be object of further investigations.

The analysis step provides amplitude, frequency and phase for each sinusoid at fixed time instants, or *frames,* here denoted by the subscript *i*.

A convenient analysis procedure for quasi harmonic sounds is described in [7] and adopted for the examples showed in this paper, even though any of the well known analysis techniques can be used as well [1] [3] [4].

3 Phase delay representation

Once the sinusoidal parameters $A_{i,k}$ $\omega_{i,k}$ $\theta_{i,k}$ have been extracted from the signal, partial phases are transformed into phase delays:

$$\tau_{i,k} = \frac{\theta_{i,k}}{\omega_{i,k}} \qquad (3)$$

Phase delays (unlike phases) are homogeneous quantities, that is phase delays belonging to two different partials can be compared. Actually, they can be interpreted as the temporal distance (in seconds) between the frame center and the nearest partial maximum (see *Figure 1*).

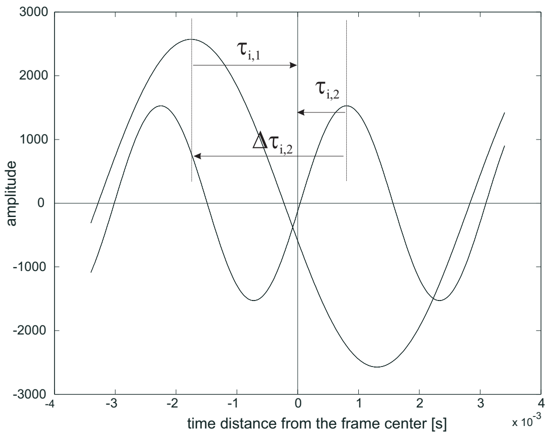

Figure 1. Illustration of the relation between phase delays $\tau_{i,k}$ and *relative phase delays* $\Delta\tau_{i,k}$. The origin of the time axis corresponds to the center of the analysis frame.

The time shifts between partials, along with amplitudes and frequencies provide a complete representation of the waveform. Equivalently, a waveform can be locally characterized by referring each phase delay to the phase delay of the first partial, by defining the *relative phase delays (rpds)*:

$$\Delta\tau_{i,k} = \tau_{i,k} - \tau_{i,1} \qquad (4)$$

The vector

$$\boldsymbol{\Delta\tau_i} = \{\Delta\tau_{i,k,}\} \qquad k=2,\dots,N \qquad (5)$$

generalizes the waveform description, as it eliminates the dependence on the phase (delay) of the first partial. In other words, given an arbitrary value for the phase of the fundamental, the original waveform can be built *on* it just by adding the relative phase delays. The equation for calculating the partial (wrapped) phases from a (modified) first partial phase and *rpds* is therefore:

$$\theta^*_{i,k} = \mathrm{mod}\left\{\left(\frac{\theta^*_{i,1}}{\omega_{i,1}} + \Delta\tau_{i,k}\right)\omega_{i,k}, 2\pi\right\} \qquad k = 2,\dots,N \quad (6)$$

where $\theta^*_{i,k}$ is an arbitrary real value. The only difference between the original and the modified waveform, characterized by ($A_{i,k}$,$\omega_{i,k}$,$\theta_{i,k}$) and ($A_{i,k}$,$\omega_{i,k}$,$\theta^*_{i,k}$) respectively, will result in a time shift.

Here we want to emphasize that *rpds* plus the first partial phase convey the same information as the phases. For this reason, hereafter, the expression *sinusoidal description* will refer, without distinction, to the amplitudes + frequencies + phases or to the amplitudes + frequencies + *rpds* + fundamental phase representations.

4 Time stretching algorithm

The traditional way of performing time stretching by sinusoidal models is to resample the frequency and amplitude tracks at higher or lower rates, thus speeding up or down the performance without affecting the spectral envelope. The method proposed in this paper can be considered as an extension, since, besides amplitude and frequency, also the phase, in its *rpd* form, can be interpolated /decimated. The basic idea is to use the fundamental as a 'carrier' for the upper partials: its phase is time scaled on the base of a continuity principle (see below) and interpolated in between the frames, while the phases of the upper partials are reconstructed by using (6).

4.1 Normalized Relative Phase Delays

The representation given in (4) needs to be slightly modified to be actually implemented in the algorithm. Analysis phases are wrapped, so they are defined except for an integer number of 2π. This uncertainty affects *rpd* of a partial in terms of an unknown added number of partial periods. Since we want to be able to interpolate the *rpds* between frames, we must make sure that, at least for a quasi stationary waveform, *rpds* are consistent for adjacent frames. For this purpose we define the *normalized relative phase delays (nrpds)* $\tilde{\Delta}\tau_{i,k}$ by adding to the *rpds* an integer number $M_{i,k}$ of the *k*-th partial periods so as to impose that the *nrpds* lie in the range [0, $2\pi/\omega_i$):

$$\tilde{\Delta}\tau_{i,k} = \Delta\tau_{i,k} + 2\pi\frac{M_{i,k}}{\omega_{i,k}} \qquad (7)$$

$$0 \le \tilde{\Delta}\tau_{i,k} < \frac{2\pi}{\omega_{i,k}} \Rightarrow M_{i,k} = \left\lfloor 1 - \frac{1}{2\pi}\Delta\tau_{i,k}\omega_{i,k} \right\rfloor \qquad (8)$$

where $\lfloor x \rfloor$ indicates the greatest integer below x. A plot of *nrpd* vs. partial number of a quasi stationary waveform superimposed over many successive frames is shown in *Figure 2b*. It is evident the *nrpd* coherence.

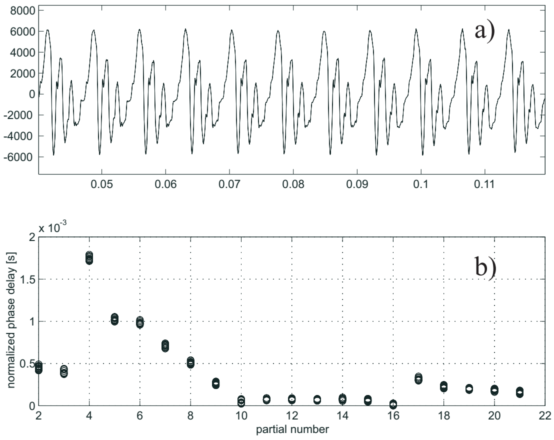

Figure 2. Plot of the normalized relative phase delays (*nrpds*) for a quasi stationary portion of voiced speech. a) Original waveform. b) Plot of the *nrpds* versus partial number for 10 successive analysis frames.

Even *nrpds* are occasionally affected by jumps, when they would cross zero or the partial period limits. However, these jumps are easily recognized and corrected at interpolation time.

4.2 First partial phase trajectory

The next step is the determination of the phase trajectory of the modified first partial. The analysis frame duration T, here supposed constant for simplicity, is scaled according to the time stretch factors ρ_i (which may change from frame to frame for piecewise-constant time varying modifications) and the modified frame locations $\hat{B}_i$ are computed as follows:

$$\hat{T}_i = \rho_i T \qquad \hat{B}_i = \sum_{l=1}^{i} \hat{T}_i \tag{9}$$

The modified fundamental phase $\hat{\theta}_{i,1}$ is evaluated at the new frame locations by the propagation formula:

$$\hat{\theta}_{i,1} = \hat{\theta}_{i-1,1} + \rho_i(\theta_{i,1} - \theta_{i-1,1}) \tag{10}$$

which is exact if ρ_i is constant within the i-th frame. All phases in (10) are intended as unwrapped. The unwrapping procedure adopted here is the one proposed by McAulay and Quatieri [1], based on the minimization of the mean square of the second derivative of the analysis phase.

Equation (6) accomplishes the requirement of waveform preservation locally, i.e. around the (modified) frame boundary, thus ensuring the so called *intraframe* phase coherence [4]. It is possible to demonstrate by straightforward algebra that, as long as the sound is (quasi) harmonic, the same formula ensure also the *interframe* coherence, that is the phase coherence across the frames for all partials, provided that the phase of the first partial is updated by (10).

At this point we have the description of the time stretched sound at the modified frame locations $\hat{B}_i$. Note that this description coincides with that of the input signal for $A_{i,k}$, $\omega_{i,k}$, $\tilde{\Delta}\tau_{i,k}$ and the only parameter that has changed is the first partial phase which is calculated through (10).

4.3 Synthesis

The output signal could be now synthesized by a variable frame length synthesis algorithm, like the classical cubic phase interpolation [1], but a constant frame rate is often more attractive when the signal has to be synthesized by other methods, namely by IFFT. Actually, by interpolating the sinusoidal representation, the output frame locations can be made arbitrary. Defined T_O as the output framing interval, the interpolation process first locates the desired n-th frame positions nT_O with respect to the $\hat{B}_i$ by searching for the indexes j and j+1 for which $\hat{B}_j \le nT_O < \hat{B}_{j+1}$. Then a linear interpolation between $\hat{B}_j$ and $\hat{B}_{j+1}$ is performed for amplitudes, frequencies and *nrpds*, while the first phase is evaluated by cubic interpolation thus ensuring the appropriate smoothness. Finally, partial phases are recovered by using (6), where $\Delta\tau_{i,k}$ are replaced by the interpolated $\tilde{\Delta}\tau_{i,k}$.

5 Extension to pitch shifting

A straightforward extension of the above process to frequency scaling comes from the observation that, provided the frequency scaling factor β_i is constant over the duration of the frame, the phase variation induced by the frequency scaling on a sinusoid is equivalent to that produced by a time stretching with the same expansion factor. Thus, joint time stretching and pitch shifting is obtained by substituting ρ_i with $\rho_i\beta_i$ in equation (10) and $\omega_{i,k}$ with $\beta_i\omega_{i,k}$ in (6) but leaving (9) untouched. With these substitution, the new synthesis equations become:

$$\hat{\theta}_{i,1} = \hat{\theta}_{i-1,1} + \rho_i\beta_i(\theta_{i,1} - \theta_{i-1,1}) \tag{11}$$

$$\hat{\theta}_{i,k} = \mathrm{mod}\left\{\left(\frac{\hat{\theta}_{i,1}}{\beta_i\omega_{i,1}} + \tilde{\Delta}\tau_{i,k}\right)\beta_i\omega_{i,k}, 2\pi\right\} \qquad k = 2,\ldots,N \tag{12}$$

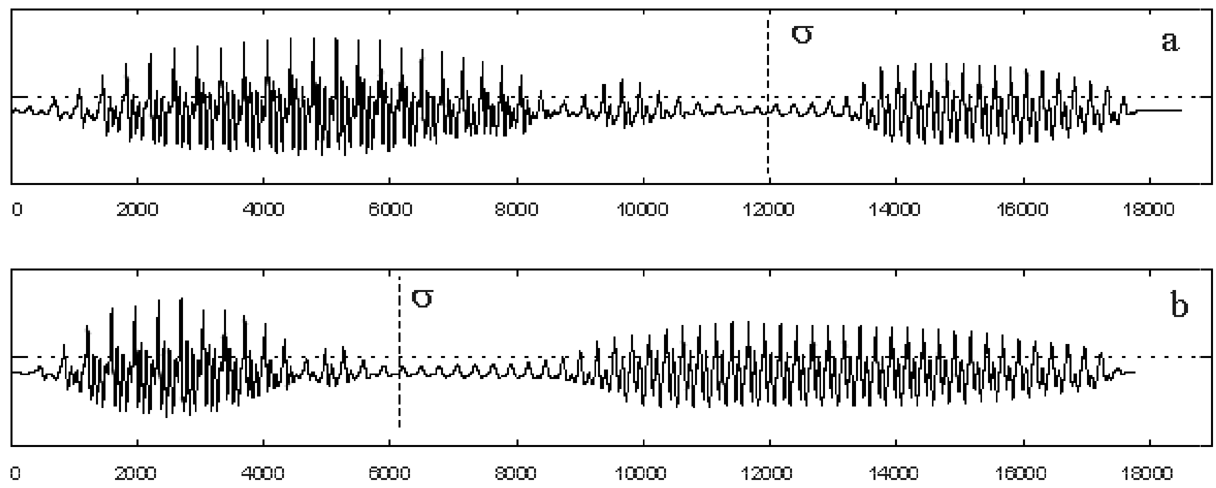

Figure 3. Time stretching with piecewise constant scale factor. a) Original signal. b) Modified signal. Time compression from beginning to σ ($\rho = 0.5$) and expansion from σ to the end ($\rho = 2$).

Finally, in order to keep the original formant structure, amplitudes are adjusted by linear interpolation on the original spectral envelope. It can be shown that *interframe* coherence is still preserved.

6 Results

The algorithm has been tested on various sounds with harmonic or quasi harmonic structure, including singing, violin and clarinet. High quality results could be obtained even for very large time stretching factors, up to 30 and over. In *Figure 3* a the Italian word 'verde' has been modified by the time stretching algorithm with piecewise constant scale factor. Joined time scale and pitch modifications were tested as well, showing very good results.

7 Conclusions

A system for time stretching and pitch modification of quasi harmonic sounds has been presented. The system, developed for the sinusoidal modeling framework, produces very high quality results in terms of perceived sound and keeps the computational load low, just slightly over the classical 'magnitude only reconstruction' method, as the basic difference is the inclusion of the phase contribution by *relative phase delays*. Compared to other methods [2], the presented system is simpler (it does not require signal deconvolution nor complex interpolation of the spectral evelope) and more robust to pitch inaccuracies, since the pitch is less important for the algorithm.

Developments of the system are being considered for processing and incorporation of the residual, here neglected, and the generalization of the *relative phase delay model* to the individual modification of the partials.

Acknowledgements

This work has been supported by Telecom Italia S.p.A, under the research contract 'Cantieri Multimediali'.

References

[1] R. J. McAulay, T. F. Quatieri "Speech Analysis/ Synthesis based on a sinusoidal representation," IEEE Trans. ASSP vol. 34 No. 4 August 1986, pp.744- 754.

[2] T. F. Quatieri, R. J. McAulay, "Shape Invariant Time-scale and Pitch modification of Speech," IEEE Trans. On Sig. Proc. vol. 40 No. 3 March 1992, pp. 497- 510.

[3] X. Serra "Musical Sound Modeling with sinusoid plus noise," in *Musical Signal Processing* Ed. by C. Roads, S. T. Pope, A. Piccialli and G. De Poli, Swets and Zeitlinger Publ. pp. 91-122, 1997.

[4] E. B. George, M. J. T. Smith "Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model," IEEE Trans. ASSP vol. 5 No. 4 September 1997, pp.389-406.

[5] M. Dolson "The Phase Vocoder: A Tutorial," Computer Music Journal, vol. 10 No. 4 Winter 1986, pp.14-27.

[6] M. P. Pollard et. al. "Shape invariant pitch and time scale modification of speech by variable order phase interpolation," Proceedings of the IEEE ICASSP97, vol.2, pp.919-922.

[7] R. Di Federico, G. Borin, "An improved pitch synchronous sinusoidal analysis - synthesis method for voice and quasi harmonic sounds," XII Colloquium on Musical Informatics, Gorizia (Italy), 1998

[8] L. R. Rabiner, R. W. Schafer, *Digital processing of Speech Signals*. Englewood Cliffs: NJ: Prentice-Hall, 1978.