AN IMPROVED HYBRID TIME-FREQUENCY ALGORITHM FOR TIME-SCALE MODIFICATION OF SPEECH/AUDIO SIGNALS

Cristian NEGRESCU, Amelia CIOBANU, Dragoş BURILEANU, Dumitru STANOMIR

Faculty of Electronics, Communications and Information Technology, University "Politehnica" of Bucharest

Corresponding author: Cristian NEGRESCU
P.O. Box 35-145, Bucharest, 050461, Romania
negrescu@elcom.pub.ro

The paper presents a new hybrid algorithm for time-scale modification, based on the Knox, Bailey and Stewart solution (KBS-TSM). The proposed algorithm improves the segmentation and concatenation methods and uses a more elegant spectral analysis involving instantaneous frequency (IF) spectrogram. The efficiency of the new algorithm is increased by including the IF attractors as a better estimation for the frequencies of the partials. Finally, a better balance between temporal and frequency resolution is obtained by decoupling the lengths of analysis and synthesis frames for the sinusoidal modelling.

Key words: Time-scale modification, sinusoidal model, instantaneous frequency, IF attractors.

1. INTRODUCTION

A time-scale modification algorithm for speech or audio signals transforms a known input signal into a replica of it with another duration, but retaining all other perceptual characteristics of the original one. Browsing a large number of techniques and implementations, two basic approaches become prominent – time domain and frequency domain [1]. Time domain algorithms, such as the very popular synchronized overlap-add (SOLA), are simple, efficient and offer good results when highly correlated, quasi-periodic content is involved [2]. On the other side, the frequency domain algorithms can perform time-scaling of signals with more complex spectral composition but are more expensive in terms of complexity and without special precautions, these algorithms can introduce reverberant or phase artefacts. A large number of frequency domain methods are based on the original vocoder idea [3] enriched with various improvements [4]. Typically, for speech signals involving only one speaker at a time, one prefers time domain algorithms, while the frequency domain methods are more used for complex sounds, such as multi-pitch speech signals or polyphonic music. Another limitation of time domain algorithms is given by the intrinsic working mode (elimination or repetition of some parts from the original signal). Consequently, such algorithms can be used only for small time-scale factors, up to +/-30% for single-pitch speech signals and no more than +/-15% for polyphonic audio signals. Beyond these stretching ratios, frequency domain algorithms remain favourite.

We intend to address applications where complex signals (i.e., pure speech, pure instrumental music, singing voice, multi-pitch speech signals, mono/multi singing voices with musical background) are reasonably well time-scaled involving compression ratios up to 40% or extensions no more than 50%, but without paying the full price of implementing frequency domain algorithms. For such applications, a hybrid solution can combine the simplicity of time domain algorithms with the high quality and flexibility of frequency domain algorithms. In the area of single-pitch speech signals the time-scale modification algorithm proposed by Knox, Bailey and Stewart algorithm [5] (named in the following KBS-TSM) represents a good choice. It is a well balanced hybrid algorithm which keeps the computational effort to a low level and minimizes the distortions associated with time domain algorithms using a local spectral analysis-synthesis technique. In the paper we propose a new algorithm, which improves significantly the performances of KBS-TSM and extends its ability to operate with general audio signals.

The paper is structured as follows. In chapter 2 an overview of the SOLA time domain algorithms for time-scale modification is completed with a review of associated distortions reported in the literature. The attention is focused mainly on the artefacts introduced at the edge of the frames, the main causes are identified and the factors that play important roles in their perceptions are revealed. The 3-rd chapter contains a short and general presentation of the proposed algorithm's architecture. The next two chapters represent the major contribution of this paper. Chapter 4 deals with the segment alignment and concatenation algorithm. Chapter 5 spans the entire area belonging to the spectral smoothing algorithm. It begins with an overview. The STFT time-frequency analysis used in common algorithms (including [5]) is replaced with a more refined instantaneous frequency (IF) analysis. We present the IF concept for analog processing and we derive the discrete-time implementation used in our application. The simple peak extraction routine (a method largely met in literature), that usually accompanies the sinusoidal model is substituted with a more accurate and efficient solution – the IF attractors. Finally, we proved that a flexible selection of the analysis and synthesis frames leads to a better control of the time and frequency resolution. Chapter 6 and 7 are reserved for experimental results and conclusions.

2. SYNCHRONIZED OVERLAP-ADD TECHNIQUE. ARTEFACTS

The core of time-scale modification algorithm presented in this paper is the well-known synchronized overlap-add (SOLA) technique. Basically, the SOLA technique operates by repeating (time expansion) or removing (time compression) appropriate segments of the input. In respect to "S" from the SOLA acronym, "appropriate" refers to the duration of the repeated or removed segments, which should be a multiple of the local pitch period (if it exists). Despite the described temporal "synchronization", at the edge of segments, temporal discontinuities are typically large and consequently perceptually objectionable distortions will be introduced in the time-scaled output signal.

The traditional solution offered by SOLA involves time domain averaging in the neighbourhood of the edges by weighted overlap and add technique. First, the original signal is segmented in overlapping frames with given length, $Nseg$ and frame advance, $NsegA$ (in samples). Then, to obtain the time-scaled output signal, the successive frames are placed in the same order but with another time advance, $\alpha \cdot NsegA + Ncor$, where α is the desired time-scale factor and $Ncor$ is an additional offset which is responsible for maintaining the pitch synchronization between each two successive overlapping frames (see Figure 1).

It is obvious that constant SOLA parameters simplify the implementation, but due to the nonstationarity of speech/audio signals, this constraint cannot be imposed for the whole set of parameters. As an example, if one imposes constant values for the frame length $Nseg$ and advance $NsegA$, the additional offset $Ncor$ will vary from one segment to another. Reversed, in some of the first versions of SOLA for speech signals, to avoid the computing of $Ncor$ and to simplify the overlap-add synthesis process, using a pitch detection routine followed by introducing pitch markers on the original signal, a variable and pitch synchronous primary segmentation is imposed.

Today the typical SOLA approach deals evenly with constant or adaptive segment lengths (chosen for other reasons than to obtain directly the pitch synchronization). Nevertheless, if the original segments have constant or variable length, a correlation based procedure is responsible for computing optimized $Ncor$ values for each frame. This method has good results if, the input signal is a single-pitch signal locally (valid assumption for speech sounds produced by a single speaker or for very simple musical excerpts).

SOLA, as a time domain algorithm for time-scale modification, is an attractive solution because it is simple and has the potential to produce high quality output signal. These advantages come from the fact that some parts of the input signal are copied directly into the output signal, with no additional processing. Therefore, high quality results are expected for time-scale factors close to 1, since only small parts of the signal are missing or are being repeated.

The major problems of this method for reasonable α factors are caused by discontinuities at the edges of the frames. These discontinuities refer to all the characteristics of a possible spectral component: magnitude, frequency and phase. Partially, the discontinuities are solved by SOLA's classical solution – the temporal averaging, but this method itself generates audible artifacts. Indeed, by temporal averaging, the magnitude differences at the edges of two consecutive segments are hidden, but the frequency and phase

discontinuities are not. Due to the phase mismatch, in the area where the temporal averaging is performed, some tonal components are attenuated or even discarded, the global loudness is decreased and some "beating effects" are perceived. In addition, if some of the cancelled tonal components represent harmonics (over the 2000 Hz) of the pitch frequency, the sound becomes "dull" [6].

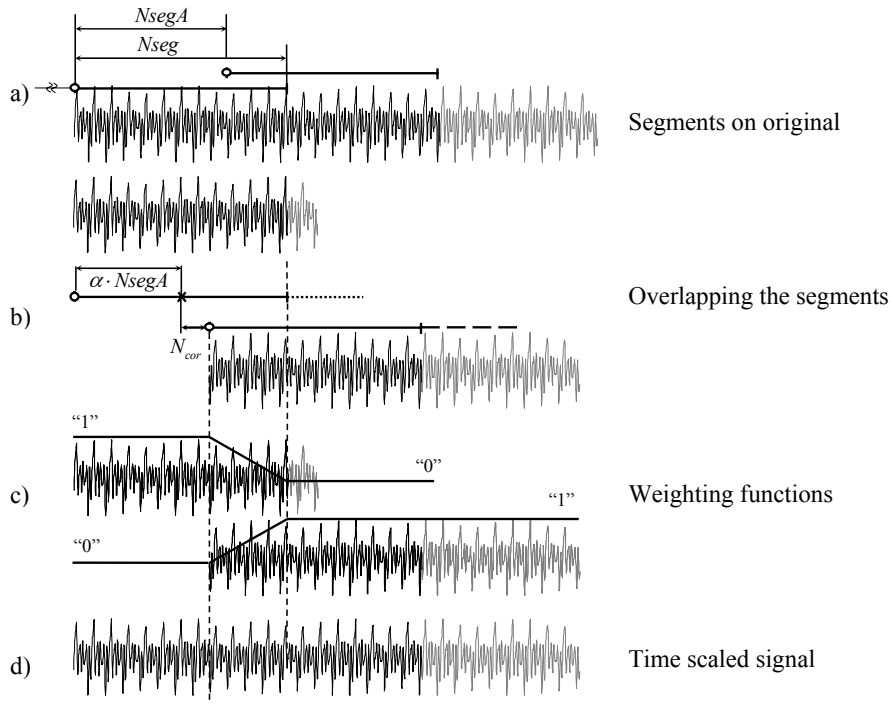

Figure1. Classical SOLA approach for $\alpha < 1$.

Having in mind the time-frequency representation of the signal (the spectrogram), one can remark that for SOLA, by adding or removing some parts from the original signal, the temporal (horizontal) synchronization is broken, but due to the direct transfer (copy) of the input segment into the synthesized output, the frequency (vertical) synchronization between different spectral coefficients tends to remain unaffected.

Regarding the edge phase discontinuities, SOLA is trying to solve them when computing $Ncor$. Taking into account the physical meaning of the autocorrelation function it becomes clear that SOLA is trying to match (globally) the phase of the spectral components that gives the specific temporal pattern of the pitch. Therefore, one can say that the horizontal phase synchronization is somehow performed, but only for the most large (and probably) relevant spectral tonal components that define the pitch. The horizontal synchronization via correlation requires an acceptable periodic nature of the input signal. Therefore, two important consequences emerge directly. First, when two segments with more spectral complex signals such as polyphonic music or multiple speakers' speech are compared, it is unlikely to find a high level of similarity between them, audible artifacts will be introduced and the signal becomes "rough". Second, when the compared signals have different nature (for example when a voiced part of a speech signal of a quasi-periodic music excerpts is overlapped and added with an unvoiced or a musical transient region), again the correlation between them is small and OLA approach contributes to the dullness of the output signal.

According to [5], there is another factor that plays a role in the perceived roughness of the delivered signal, one that is not typically addressed in SOLA algorithms or extensions of them. Even for purely voiced signals, where due to the high level of correlation, one can expect a good behavior of classical SOLA, the roughness of the signal becomes noticeable if the dilatation becomes large. The cause of this effect consists in the steep modification of frequencies (significant mismatch in the frequency domain) because adjacent subsegments in the output signal come from more disparate region of the input signal [5].

3. THE PROPOSED ALGORITHM

For speech signals an attractive and simple hybrid (time-frequency) solution was proposed by Knox, Bailey and Stewart [5]. In the following, we shall refer to this algorithm as the KBS-TSM algorithm. Our proposed algorithm represents an improved version of the KBS-TSM algorithm. Its diagram block is presented in Figure 2.

The basic segmentation idea from SOLA was kept, but due to the number of the subsidiary problems mentioned before, they renounced the time domain average. Instead, the edge discontinuities are solved using a frequency domain smoothing. So, in the first stage, two successive output segments will be concatenated by joining edges, without any overlapping. Before the concatenation, the segments will be pitch-synchronized using the correlation function. This synchronization will be performed in the vicinity of the concatenation point, but only if both segments contain voiced signals.

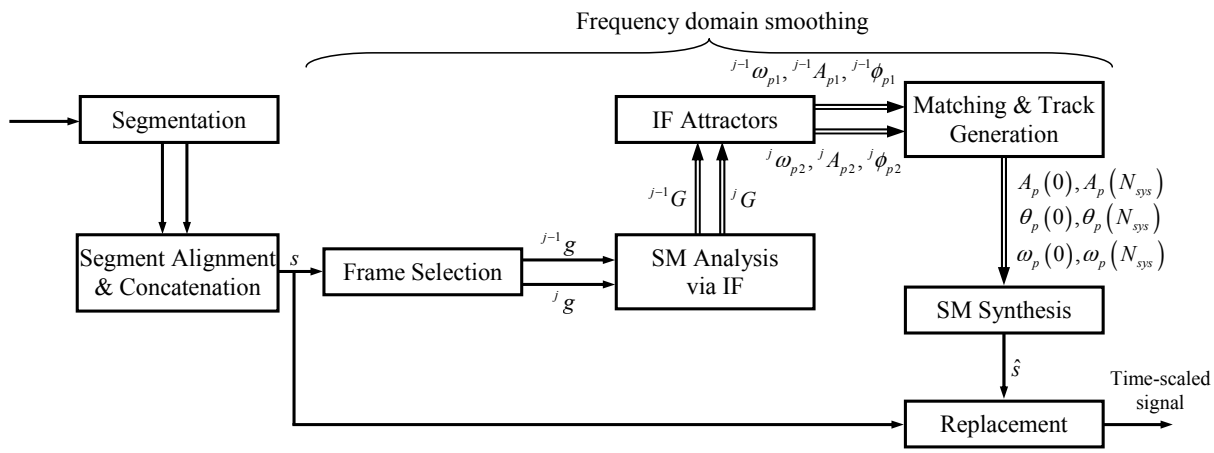

Figure 2. Diagram block for the proposed algorithm.

In the proposed algorithm the frequency domain smoothing is performed via an Analysis – IF Attractors Matching – Synthesis procedure. Based on the McAulay/Quatiery sinusoidal model (SM) [7], two signal analyses are initiated, one for each side of the concatenation point. Using only the frequency information of the spectral components, possible correspondences between nearest spectral peaks are revealed. Then, based on the same McAulay/Quatiery model, the identified IF attractors are linked across segment concatenation point and a resynthesized signal is produced. This new signal, able to perform a timbre morphing between concatenated segments, replaces the original signal spanning the concatenation location. Further details can be found in §4 and §5.

4. SEGMENT ALIGNMENT AND CONCATENATION ALGORITHM

The algorithm starts with a primary disjoint segmentation ($NsegA = Nseg$). This allows defining the cutting locations on the original signal and subsequently, the joining locations according to the TSM ratio. Focusing the attention on the jth cutting point only, two waveform segments (let it be denoted with indexes $j-1$ and j) are selected from the original signal. The first segment ($j-1$) ends with the sample placed at location indicated by the TSM required ratio, while first sample of the current segment (j) starts with the sample corresponding to the jth cutting point. At the beginning, the segments are conceptually superposed, the current segment j is placed with its start over the sample (from the segment $j-1$) corresponding to the position indicated by the TSM ratio.

We intend to perform the pitch synchronization (computing N_{cor}) by searching the best match of a large enough area located at the beginning of the segment j with the surrounding of the TSM ratio location

from the segment $j-1$. To complete this task he segment $j-1$ is extended over the position indicated by the TSM ratio. To find the optimum position of the second segment (j), two frames (${}^{j-1}f(n)$ and ${}^{j}f(n)$) are selected (see Figure 3).

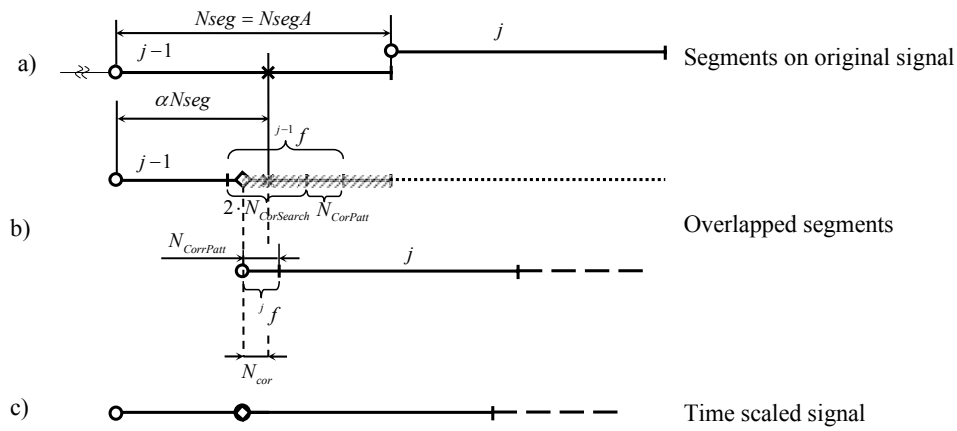

Figure 3. Segmentation and frame selection in the proposed algorithm.

The frame ${}^{j}f$ contains the first $N_{CorPatt}$ samples of the segment j and represents the correlation pattern. The second frame, ${}^{j-1}f$, contains $2N_{CorSearch}+N_{CorPatt}$ samples where $N_{CorSearch}$ represents the correlation depth (the maximum displacement on each part of the TSM ratio location for which the search is performed). The origin of the time for the frame ${}^{j-1}f$ was chosen to have ${}^{j-1}f(0)$ as the sample corresponding to the TSM ratio location.

For the correlation estimator we define

$$ {}^{j-1,j}\rho(\lambda)=\frac{\sum_{n=0}^{N_{CorPatt}-1} {}^{j}f(n)\ {}^{j-1}f(n+\lambda)}{\sum_{n=0}^{N_{CorPatt}-1}\left|{}^{j}f(n)\right|^{2}\sum_{n=0}^{N_{CorPatt}-1}\left|{}^{j-1}f(n+\lambda)\right|^{2}}, \tag{1} $$

where

$$ \lambda=-N_{CorSeach},\ ...,\ -1,0,1,\ ...,\ N_{CorSeach}\ . \tag{2} $$

The optimum displacement N_{cor} is the value of the lag which corresponds to the maximum of the cross-correlation coefficient, i.e.

$$ {}^{j}N_{cor}=\left\{\lambda_{opt}\middle|\forall\lambda=\{-N_{CorSearch},\ ...,\ N_{CorSearch}\}\setminus\{\lambda_{opt}\},\ {}^{j-1,j}\rho(\lambda_{opt})\geq{}^{j-1,j}\rho(\lambda)\right\}. \tag{3} $$

Using (1) in conjunction with the definition of the frames, we decoupled the quality of the estimator (the correlation estimator is unbiased, energy independent and its constant dispersion is controlled via $N_{CorPatt}$) from the search area (given exclusively by the $N_{CorSearch}$ value).

To expand the applicability of the time-scale modification algorithm beyond single-pitch signals, we avoid the hard voiced/unvoiced decision via the zero-crossing measures applied for the frames. Instead, we always evaluate the correlation coefficients ${}^{j-1,j}\rho(\lambda)$ and once ${}^{j}N_{cor}$ is computed, we make the decision of actual application of ${}^{j}N_{cor}$ only if the corresponding measure of similarity (${}^{j-1,j}\rho(N_{cor})$) is above a certain threshold. If not, we consider that the similarity is too low and we set ${}^{j}N_{cor}=0$ (no additional displacement).

5. THE PROPOSED SPECTRAL SMOOTHING ALGORITHM

5.1. Overview

The signal model used by the spectral smoothing algorithm is the well-known sinusoidal model [7]. According to it, the original signal $s(n)$ can be approximated (resynthesized) by adding together the contributions of all P partials. Each partial corresponds to an harmonic controlled oscillator. Consequently, the resynthesized signal $\hat{s}(n)$ can be written as

$$\hat{s}(n) = \sum_{p=1}^{P} A_p(n) \cos\left(\omega_p(n) \cdot n \cdot Ts + \phi_p\right), \tag{4}$$

where Ts is the sampling period (in [s]), $A_p(n)$ is the instantaneous amplitude, $\omega_p(n)$ (expressed in [rad/s]) is the instantaneous frequency, and ϕ_p represents the initial phase (i.e., the phase of the cosine at the moment $n = 0$). It is assumed that the signal evolves on a duration that is significantly longer than the underlying signal's periods. The quality of the approximation is expressed via a cost function (typically least square) built on the basis of the residual (approximation error) signal.

Usually, the analysis stage for the sinusoidal model is a frame based process. On each analysis frame, the analysis stage is charged with the estimation of the amplitude, frequency and phase for the constituent partials. This estimation is achieved by minimizing the cost function for the residual. Typically, the estimation is performed by peak picking in the short-time Fourier domain (as in the present paper) [7], but it can also be realized using analysis-by-synthesis techniques that explicitly minimize a mean square error in terms of the partial's parameters [8], [9] or within the general framework of matching pursuits with a static or dynamic dictionary of signals [10], [11].

The synthesis stage connects the frame parameter model (delivered by the analysis stage) via a spectral tracking process. Starting from these tracks and using low-order polynomial interpolation (linear interpolation for amplitude and cubic interpolation for phase), the sample based control functions for harmonic oscillators are derived.

Taking into account our particular problem, a smooth connection between two locations (n_L and n_R on the time-scaled signal) is desired. To accomplish this task, two SM frames ${}^{j-1}g$ and ${}^{j}g$,

$${}^{j-1,j}g(n) = s\left(n + n_{1,2} - \frac{N_{sm}}{2}\right), \qquad n = 0,1,\ldots,N_{sm}-1, \tag{5}$$

are built [1] by the "Frame Selection" block from the Figure 2. The sinusoidal model analysis procedure (the "SM Analysis" block in Figure 2) is based on instantaneous frequencies (IF). The SM analysis is applied for each frame and it generates two sets of parameters. The IF analysis stage involves weighting the signal from each frame (once with a Hamming window with N_{sm} samples and once with the first derivative of the Hamming window – see §5.2) followed by spectral decompositions using the Discrete Fourier Transform (DFT). The analysis model based on instantaneous frequencies assumes that the IF attractors correspond to the underlying sine waves and therefore, the amplitude, frequency and phase of each tonal underlying component can be obtained by evaluating the STFT (Short Term Fourier Transport) at the frequency given by the attractors. The IF representation is explored and the location of the IF attractors are revealed (see §5.3) According to this statement, IF attractor ω_0 and its corresponding amplitude and phase from interpolated STFT (the triad $\left\{{}^{l}\omega_0, {}^{l}A_0, {}^{l}\phi_0\right\}$) define completely the partial p_0 for the frame l. Let us denote the frequencies of the IF attractors for the frame $j-1$ with ${}^{j-1}\omega_{p1}$ where $p1 = 1,2,\ldots,P_1$ and respectively, for

[1] Locations n_1 and n_2 are defined on the original signal or on the extended segments and correspond to positions n_L and n_R on the time-scaled signal (see §5.4).

the frame j with ${}^{j}\omega_{p2}$, where $p2 = 1,2,...,P_2$. Their corresponding amplitude and phases will be $\left\{ {}^{j-1}A_{p1}, {}^{j-1}\phi_{p1} \right\}_{p1=1,...,P_1}$ and $\left\{ {}^{j}A_{p2}, {}^{j}\phi_{p2} \right\}_{p2=1,...,P_2}$ (see Figure 2).

For each analysis frame, the block "IF Attractors" in Figure 2 is responsible with the identification of these parameters from the corresponding IF representation.

The "Matching and Track Generation" block analyzes the two sets of data and attempts to connect pairs of attractors (one from the set $j-1$ with one from the set j). A tracking procedure generates temporal trajectories by cross-matching the peaks from these two sets in accord with certain matching criteria. In the proposed algorithm, the criteria are derived from the original McAuley/Quatieri implementation [7]. The list of peaks from the set $j-1$ is explored and for each frequency ${}^{j-1}\omega_{p1}$ a matching frequency is searched in the second set (j). The nearest ${}^{j}\omega_{p2}$ from the set j is found and if the absolute difference between ${}^{j-1}\omega_{p1}$ and ${}^{j}\omega_{p2}$ is under a specified limit $\Delta\omega_{\max}$, an initial match is defined. Following the initial match, possible conflict situation are solved. A conflict appears when the same frequency ${}^{j}\omega_{p2}$ was matched with different frequencies from the set $j-1$. The match which corresponds to a minimum frequency difference is maintained while the other initial match is broken and the corresponding ${}^{j-1}\omega_{p1}$ remains free.

Once the frequency based proximity matching process is finished, the tracks can be generated. For each matched pair of IF attractors (${}^{j-1}\omega_{p1}$, ${}^{j}\omega_{p2}$), a new track is defined by its limits ($A_p(0), A_p(N_{sys}), \phi_p(0)$, $\phi_p(N_{sys}), \omega_p(0), \omega_p(N_{sys})$).

Between the locations corresponding to n_1 and n_2 (centers of the analysis frames ${}^{j-1}g$ and ${}^{j}g$) the signal will be synthesized from its constituent partials. The "SM Synthesis" block (see Figure 2) receives the information about the tracks and, via a controlled bank of oscillators, in accord with the sinusoidal model, generates N_{sys} samples of the synthesized signal.

$$\hat{s}(n) = \sum_{p=1}^{P} A_p(n)\cos\theta_p(n), \qquad n = 0,1,...,N_{sys-1}. \tag{6}$$

To control the oscillators, $2\times P$ discrete-time functions are required: instantaneous amplitudes $A_p(n)$ and instantaneous phases $\theta_p(n)$.

Obtaining a smooth spectral transition during the length of the synthesis frame requires amplitude and phase interpolation. To complete this task, as in [5], we use the interpolation algorithm proposed by McAuley/Quatieri [7]. The sample rate control functions for instantaneous amplitudes $A_p(n)$ are obtained using a simple linear interpolation between the measured boundaries ($A_p(0)$ and $A_p(N_{sys})$), while the instantaneous phase $\theta_p(n)$ is computed performing a constrained cubic polynomial interpolation which requires $\phi_p(0)$, $\phi_p(N_{sys}), \omega_p(0), \omega_p(N_{sys})$.

The synthesized signal replaces the corresponding frame from the time-scaled version of the original signal delivered by the "Segment alignment and concatenation" block. To smooth any remaining amplitude difference, a small overlap (□ 1ms) is performed. The "Replacement" block is responsible with these last two operations.

5.2. SM Analysis using Instantaneous Frequency

In comparison with a rectangular window, the use of the Hamming window reduces the spectral leakage, but it is known that numerous spurious spectral components are still present. As a conclusion STFT does not provide good time-frequency resolution because the spectral peaks of the tonal components are smeared by the analysis window. This phenomenon alters the frequency-amplitude-phase triad estimation

and the efficiency of the approximation of the original signal using the sinusoidal model. An interesting and efficient solution could be the instantaneous-frequency (IF) spectrogram introduced by Abe, Kobayashi and Imay [14]. In the IF spectrogram the tonal components have very sharp peaks and the harmonic structure is exposed clearly. Further, we review the IF concept related to the sinusoidal model. We also derive the actual discrete-time algorithm used in our implementation.

Let us consider an analog real signal $s(t)$ written as an additive contribution of some real partials

$$s(t)=\sum_{p=1}^{P} s_p(t), \tag{7}$$

each partial p having the explicit expression of a modulated harmonic component

$$s_p(t)=A_p(t)\cos\left(\theta_p(t)\right)\left[\sigma\left(t-t_{sp}\right)-\sigma\left(t-t_{ep}\right)\right], \tag{8}$$

where $\sigma(t)$ is the Heaviside function, $A_p(t)$ represents the instantaneous amplitude and the argument of the cosine, denoted with $\theta_p(t)$, is named instantaneous phase. One can remark that in the time domain, the existence of each partial can be limited to a certain interval, $\left[t_{sp}, t_{ep}\right]\subset\square$. Starting from $\theta_p(t)$, one can introduce the instantaneous frequency (IF) of the partial p as

$$\omega_p(t)=\frac{d}{dt}\theta_p(t). \tag{9}$$

Denoting with ϕ_p the initial phase of the partial p, the instantaneous phase becomes

$$\theta_p(t)=\int^{t}\omega_p(\tau)d\tau+\phi_p, \tag{10}$$

and the equation (8) changes into

$$s_p(t)=A_p(t)\cos\left(\int^{t}\omega_p(\tau)d\tau+\phi_p\right)\left[\sigma\left(t-t_{sp}\right)-\sigma\left(t-t_{ep}\right)\right]. \tag{11}$$

In the following we intend to determine, for each partial, the triad: instantaneous amplitude $A_p(t)$, instantaneous frequency $\omega_p(t)$ and the initial phase ϕ_p. Let us suppose that one gives us the shape of $s_p(t)$ on a sample by sample basis and asks us to find $A_p(t)$, $\omega_p(t)$ and ϕ_p. The solution of this problem is not unique i.e., there are more than one combination of parameters that satisfy the equation (11). One of these solutions can be obtained by first building the analytical signal $\tilde{s}_p(t)$ associated with $s_p(t)$.

$$\tilde{s}_p(t)=s_p(t)+j\cdot\mathbf{H}\left\{s_p(t)\right\}, \tag{12}$$

where $\mathbf{H}\{\ \}$ is the Hilbert transform. Then, the instantaneous amplitude is obtained computing the magnitude of the analytical signal

$$A_p(t)=\left|\tilde{s}_p(t)\right|, \tag{13}$$

the initial phase is obtained by evaluating the argument of the analytical signal at the initial moment,

$$\phi_p=\arg\left\{\tilde{s}_p(t)\right\}\Big|_{t=0}, \tag{14}$$

while the instantaneous frequency can be obtained by time derivation of the same argument

$$\omega_p(t)=\frac{d}{dt}\arg\left\{\tilde{s}_p(t)\right\}. \tag{15}$$

Using the presented method, the desired parameters can be computed from the time domain evolution of the partials.

The next step forward in finding an algorithm for computing the triads $\{A_p(t),\ \omega_p(t),\ \phi_p\}$ from the initial signal $x(t)$, is to obtain the decomposition (7). An elegant solution is proposed in [10] and makes use of a set of band-pass filters. Each filter is designed and controlled to track only one harmonic component, obtaining at the output the desired partial. The filters should have narrow bandwidth and linear phase characteristics. This bank of complex filters is built by modulation of a causal and real prototype low-pass filter. Let it be $w(t)$ the impulse response (weight function) of the prototype. To preserve the amplitude, $w(t)$ should satisfy the condition

$$\int_{-\infty}^{\infty} w^2(t)\,dt = 1. \tag{16}$$

The bank of band-pass filters can be efficiently implemented using STFT. The STFT of the original signal can be obtained by applying the Fourier transform to the windowed signal. Focusing our analysis on a specific moment τ, the STFT of the original signal can be interpreted as the regular Fourier transform of the signal $s(t)w(\tau - t)$

$$S(\omega,\tau) = \int_{-\infty}^{\infty} s(t)w(\tau - t)e^{-j\omega t}dt\,. \tag{17}$$

The second interpretation follows writing the previous expression for a fixed frequency Ω_n and considering the second argument, time, as a variable. This interpretation leads us to consider the time dependent Fourier representation in terms of linear filtering. Indeed, highlighting the convolution, one can write

$$S(\Omega_n,t) = \left[s(t)e^{-j\Omega_n t}\right] * w(t)\,. \tag{18}$$

For a chosen fixed frequency Ω_n, we define a new signal

$$s_{f,\Omega_n}(t) \overset{Def}{=} S(\Omega_n,t)e^{j\Omega_n t} = \left\{\left[s(t)e^{-j\Omega_n t}\right] * w(t)\right\}e^{j\Omega_n t}\,. \tag{19}$$

The equation (19) suggests the processing system presented in Figure 4a. We denote with $W(\omega)$ the Fourier transform of the real and causal window function $w(t)$. It is a simple task to show that the presented processing system is equivalent with a complex band-pass filter, having the central frequency Ω_n and transfer function $W(\omega - \Omega_n)$ (see Figure 4b). Therefore, the output signal is the desired analytical signal which allows computing the instantaneous frequency first, then amplitude and phase.

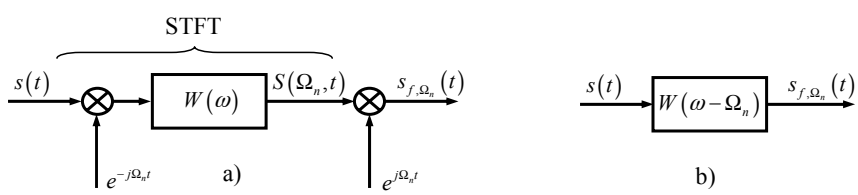

Figure 4. Band-pass filtering using STFT.

According to (15), the IF of the signal $s_{f,\Omega_n}(t)$ is

$$\omega_n(t) = \frac{d}{dt}\operatorname{arctg}\frac{I_{\Omega_n}(t)}{R_{\Omega_n}(t)}, \tag{20}$$

where

$$R_{\Omega_n}(t) = \operatorname{Re}\left\{s_{f\Omega_n}(t)\right\},\quad I_{\Omega_n}(t) = \operatorname{Im}\left\{s_{f\Omega_n}(t)\right\}. \tag{21}$$

Denoting the time derivation with a dot placed above the function, one can obtain

$$\omega_n(t) = \frac{I_{\Omega_n}(t)\dot{R}_{\Omega_n}(t) - R_{\Omega_n}(t)\dot{I}_{\Omega_n}(t)}{R^2_{\Omega_n}(t) + I^2_{\Omega_n}(t)}. \tag{22}$$

To find a convenient expression for the instantaneous frequency, the explicit derivation of sample by sample specified signals $R_{\Omega_n}(t)$ and $I_{\Omega_n}(t)$ should be avoided. The derivative of the filtered signal $s_{f,\Omega_n}(t)$ yields to

$$\frac{d}{dt}s_{f,\Omega_n}(t) = \frac{d}{dt}\left[S(\Omega_n,t)e^{j\Omega_n t}\right] = e^{j\Omega_n t}\dot{S}(\Omega_n,t) + j\Omega_n S(\Omega_n,t)e^{j\Omega_n t}. \tag{23}$$

But,

$$\dot{S}(\Omega_n,t) = \frac{d}{dt}\left\{\left[s(t)e^{-j\Omega_n t}\right] * w(t)\right\} = \left[s(t)e^{-j\Omega_n t}\right] * \dot{w}(t). \tag{24}$$

According to (24), $\dot{S}(\Omega_n,t)$ can be considered as the result of applying STFT to the same original signal $s(t)$, but the original window $w(t)$ is replaced with its derivative $\dot{w}(t)$. Due to the fact that usually $w(t)$ has a simple expression, $\dot{w}(t)$ can be precomputed using analytical derivation, without approximation errors. Finally, $\dot{R}_{\Omega_n}(t)$ and $\dot{I}_{\Omega_n}(t)$ requested by (22) become

$$\dot{R}_{\Omega_n}(t) = \mathrm{Re}\left\{e^{j\Omega_n t}\left[\dot{S}(\Omega_n,t) + j\Omega_n S(\Omega_n,t)\right]\right\}, \tag{25}$$

and respectively

$$\dot{I}_{\Omega_n}(t) = \mathrm{Im}\left\{e^{j\Omega_n t}\left[\dot{S}(\Omega_n,t) + j\Omega_n S(\Omega_n,t)\right]\right\}. \tag{26}$$

and the IF $\omega_n(t)$ can now be computed.

Let us focus our attention on the amplitude of the underlying tonal component with instantaneous frequency $\omega_n(t)$, at a fixed moment of time t. Taking into account the previous remarks on the estimation accuracy and considering the amplitude spectral density available via STFT, one can write[2]

$$\overline{A}_n(t) = \lim_{\Delta\omega\to 0}\left\{\left[\sigma(\omega-\omega_n+\Delta\omega) - \sigma(\omega-\omega_n-\Delta\omega)\right] * s_{f,\omega}(t)\right\} = \lim_{\Delta\omega\to 0}\int_{\omega_n(t)-\Delta\omega}^{\omega_n(t)+\Delta\omega} S(\omega,t)e^{j\omega t}d\omega. \tag{27}$$

which corresponds to an ideal integration of the spectral density over an infinitesimal interval around frequency $\omega_n(t)$. Finally, the instantaneous amplitude and the initial phase are

$$A_n(t) = \left|\overline{A}_n(t)\right|, \quad \phi_n(t) = \arg\left\{\overline{A}_n(t)\right\} \tag{28}$$

and, become

$$A_n(t) = \left|\lim_{\Delta\omega\to 0}\int_{\omega_n(t)-\Delta\omega}^{\omega_n(t)+\Delta\omega} S(\omega,t)e^{j\omega t}d\omega\right| \square \lim_{\Delta\omega\to 0}\int_{\omega_n(t)-\Delta\omega}^{\omega_n(t)+\Delta\omega}\left|S(\omega,t)\right|d\omega \tag{29}$$

and, respective,

$$\phi_n(t) = \arg\left\{\lim_{\Delta\omega\to 0}\int_{\omega_n(t)-\Delta\omega}^{\omega_n(t)+\Delta\omega} S(\omega,t)e^{j\omega t}d\omega\right\}. \tag{30}$$

---

[2] Here t is a constant and ω is a variable, therefore $s_{f,\omega}(t)$ is a function in frequency. Also, one can remark that $\left|s_{f,\omega}(t)\right| = \left|S(\omega,t)\right|$.

A representation of $A_n(t)$ (computed according to (29)) versus $\omega_n(t)$ is called IF amplitude spectrum [15] while the representation of $\phi_n(t)$ versus the same $\omega_n(t)$ is called IF phase spectrum.

We intend to apply the process described above for discrete time, once for each frame ${}^{j-1,j}g(n)$ (see equation (5)). The length of each frame is N_{sm} and is equal with the length of the STFT window. The discrete time starts with 0 for each frame and ends at $N_{sm}-1$. Consequently, the (continuous time) moment of interest in computing STFT is $t = N_{sm}/Fs$. The central frequencies for the band-pass filters will be $\Omega_n = n2\pi Fs/N_{sm} = n\omega_s/N_{sm}$ and the quantity $e^{j\Omega_n t}\Big|_{t=N_{sm}/Fs}$ becomes 1. The STFT will be actually computed via DFT/FFT. The whole process of evaluating the instantaneous frequencies is presented in Figure 5

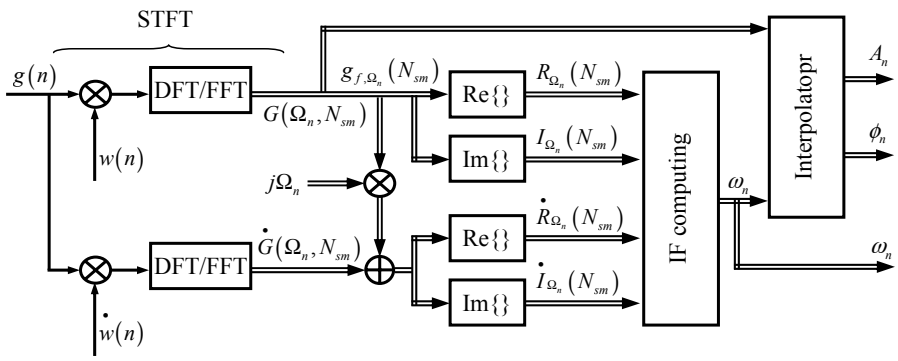

Figure 5. Diagram block for discrete time IF computing.

Because the discrete IF computing algorithm is a block algorithm, it delivers only one value of instantaneous frequency for the whole input frame. This value can be considered as the best expectation for the true instantaneous frequency in the middle of the frame, $N_{sm}/2$ (at the discrete-time moment corresponding to the symmetry point of the window). A similar remark is also true for the complex amplitude, but in this case the direct application of (27) is impossible because the STFT is available only for discrete frequencies. If the instantaneous frequency estimated before is exactly one of these discrete frequencies (let us assume to be Ω_{n_0}) the instantaneous complex amplitude is $G(\Omega_{n_0}, N_{sm})$. If the instantaneous frequency is placed between two available spectral components ($\Omega_{n_0} < \omega_n < \Omega_{n1}$) then $\overline{A}_n$ can be computed performing an interpolation of the nearest spectral components. In our implementation we used a sinc-type interpolation over $G(\Omega_{n_0}, N_{sm})$ and $G(\Omega_{n_1}, N_{sm})$, but also a simple linear interpolation can be considered an acceptable solution. Once the complex amplitude $\overline{A}_n$ is obtained, applying (28) one can compute directly the instantaneous amplitude A_n and the initial phase ϕ_n.

As a conclusion, the IF amplitude and phase spectra are based on the instantaneous frequency defined as a function of frequency and time derived from the STFT spectrogram. The IF amplitude and phase spectra are obtained by projecting the STFT amplitude spectrum onto the IF axis.

5.3. The Estimation of IF Attractors

Using the instantaneous frequencies and subsequent IF amplitude and phase spectrograms instead of the STFT allows accurate estimation of the partials parameters and therefore, the residual signal can be reduced and the smooth timber morphing area can be connected precisely with the rest of the signal (this connection is performed by the “Replacement” block in Figure 2).

We avoid the spurious spectral peaks and we increase the accuracy of partials identification replacing the simple spectra peak extraction with the construction of IF attractors. To understand this concept, let us

consider that the original analog signal contains a pure sinusoidal component with frequency ω_0. In the spectral analysis performed with STFT the components are localized around the ω_0 and are spread over a certain bandwidth depending on the analysis window (w). Applying the procedure described in the previous paragraph, if the frequency Ω_n is sufficiently close to ω_0, one obtains $\omega_n(t) = \omega_0$. According to [16] an IF attractor is the frequency Ω_n that satisfies two conditions. The first one is

$$\mu(\Omega_n, t) = 0, \tag{31}$$

where

$$\mu(\Omega_n, t) = \omega_n(t) - \Omega_n, \tag{32}$$

while the second one refers to the partial derivative of $\mu(\Omega_n, t)$ in respect to Ω_n

$$\frac{\partial \mu(\Omega_n, t)}{\partial \Omega_n} = \frac{\partial \omega_n(t)}{\partial \Omega_n} - 1 < 0. \tag{33}$$

To understand the reason of these conditions, let us consider that the original signal contains two significant tonal components with frequencies ω_0 and ω_1 added to a small wideband noise. Also, we assume that $\omega_1 > \omega_0$ and there is no (significant) interference between spectral components of the tones.

Considering Ω_n as a variable, we intend to explore, in small steps, the whole frequency domain. One can remark that if Ω_n is placed before and far enough from the first tonal component (ω_0), only the noise components placed around Ω_n (inside the bandwidth determined by the window function) affects the estimated instantaneous frequency $\omega_n(t)$.

Due to the particular shape of the transfer function of the bandpass filter, the central component is enhanced, $\omega_n(t)$ tends to be close to Ω_n, the slope $\partial \omega_n(t) / \partial \Omega_n \to 1$ and as a result,

$$\frac{\partial \mu(\Omega_n, t)}{\partial \Omega_n} \to 0. \tag{34}$$

When Ω_n increases and approaches to ω_0, the spectral components of the first tonal signal just enter inside the bandwidth of the band-pass filter. The contribution of the spectral components of the first tonal signal has an almost vertical slope (a steep increase) becoming significant.

Therefore, the instantaneous frequency is determined correctly ($\omega_n(t) = \omega_0$). In this area a small positive increment for Ω_n generates a larger positive modification of $\omega_n(t)$, and thus, the slope of $\mu(\Omega_n, t)$ as a function of Ω_n turns positive,

$$\frac{\partial \mu(\Omega_n, t)}{\partial \Omega_n} > 0. \tag{35}$$

Further increase of Ω_n up to ω_0 and even over a little area beyond (to have enough significant components for the first tonal signal inside the bandwidth), locks the instantaneous frequency on ω_0. Consequently, $\omega_n(t)$ becomes almost independent of Ω_n, $\partial \omega_n(t) / \partial \Omega_n \to 0$ and the slope becomes negative, close to -1

$$\frac{\partial \mu(\Omega_n, t)}{\partial \Omega_n} \to -1. \tag{36}$$

When additional increasing of Ω_n is performed, the bandpass filter selects only the noise components and the result is similar with the one obtained when $\Omega_n << \omega_0$ (see (34)). The entire presented process is repeated when Ω_n becomes closer, reaches or overpasses the tonal component placed at ω_1.

Taking all these aspects into account, it becomes clear that the IF attractors defined according to (31) and (33) are robust estimators for the instantaneous frequencies of the underlying tonal components. Comparing with peak extraction applied to STFT spectrogram, the representation of IF attractors is a more accurate estimator. It eliminates a significant number of false peaks introduced by the analysis window, therefore it decreases the number of partials considered by the synthesis algorithm. An important remark reveals the fact that the identification of the instantaneous frequencies of the partials as IF attractors is based only on the IF spectrum. The information regarding the amplitudes and phases are not requested.

In our application, for computing the IF estimators we can access the instantaneous frequency ω_n only for certain discrete frequencies ($\Omega_n = n2\pi Fs / N_{sm} = n\omega_s / N_{sm}$). We start with evaluation of $\mu(\Omega_n)$ for all available Ω_n. Next, using linear interpolation, we find the zero crossing values of $\mu(\Omega_n)$ and for each found zero, we check the condition (33) using finite differences. We keep as IF attractors only the interpolated frequencies which satisfy the condition (33). Finally, their amplitude and instantaneous phase are obtained via interpolation.

5.4. Analysis/Synthesis Frame Selection

As stated before, we intend to apply the sinusoidal model, which is a frame-based process, around the concatenation point and resynthesize the signal in that area. In order to complete this task we need to build the analysis and synthesis frames (the "Frame Selection" block in Figure 2 is responsible with this task).

Frame selection is done as follows. We explore the area surrounding the concatenation point and choose the points n_L and n_R between which we wish to resynthesize the signal (see Fig. 6 c). The number of samples between the selected points (N_{sys}) will determine the length of the synthesis frame. Then, on the segments prior to concatenation we identify the selected points and place the analysis frames centered on these points (see Figures 6b and 8c). Depending on the length of the synthesis frame, the analysis frames can be disjoint or not.

An obvious improvement is the fact that the analysis and synthesis are no longer coupled as in [5]. One can remark that a longer analysis frame does not impose a longer synthesis frame anymore, so the quality of spectral estimation and the timbre morphing duration can be now independent. In our application we set the analysis frames four times larger than the synthesis frames $N_{sm} = N_{sys} * 4$.

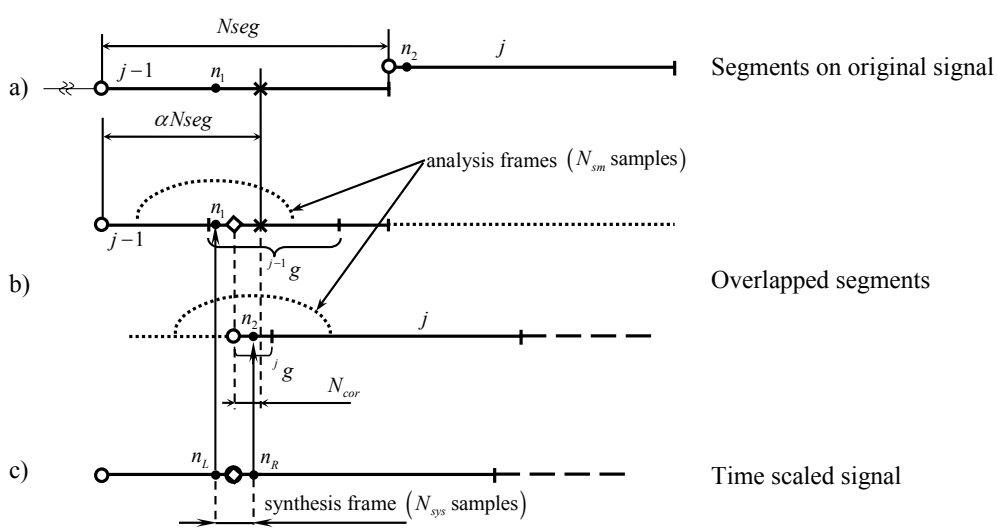

Figure 6. Frame selection in the proposed algorithm.

6. EXPERIMENTAL RESULTS

We implement the proposed algorithm as a Matlab application. The software application is controlled via a friendly GUI (see Figure 7). All the parameters of the proposed algorithm can be modified. A list of the parameters is given in Table 1.

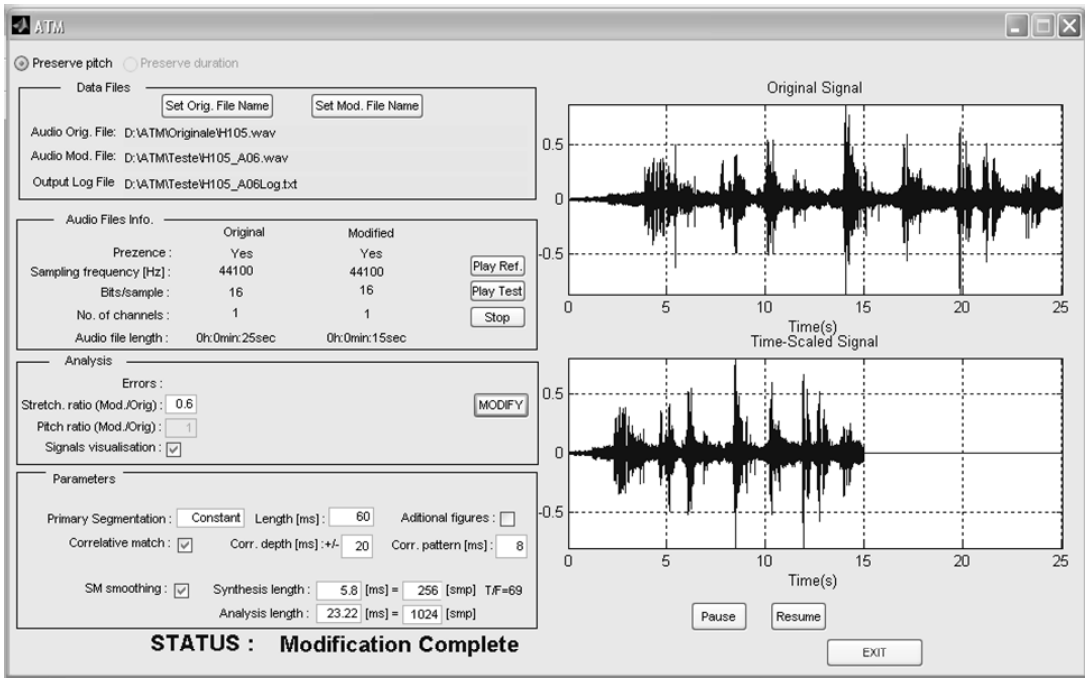

Figure 7. GUI of the software time-scale application.

Table 1. Parameters

Name	Notation	Description	Reference
TSM ratio	α	TSM ratio	See §2, §4
Primary segment	"Constant"	Method for segmentation	See §2
Length	N_{seg}	Segment length	See §2, §4
Correlative match	On/off	Enables/Disables correlative overlapping	See §2, §4
Correlation depth	$N_{CorSearch}$	Searching area for correlative concatenation	See §4
Correlation pattern	$N_{CorPatt}$	Frame length ${}^{j}f$	See §4
SM smoothing	On/off	Enables/Disables the spectral morphing algorithm	See §5
Window	w	Window function used in SM ("Hamming", "Blackman Harris")	See §5.2
Synthesis length	N_{sys}	Length of the SM synthesis frame ($n_R - n_L$)	See §5.4
Analysis length	N_{sm}	Length of the SM analysis frame	See §5.4

For tests we have selected 19 high quality monophonic signals covering different classes and genres:

- pure speech, clean background, single speaker, male – 2 signals,
- pure speech, clean background, single speaker, female – 2 signals,
- pure speech, clean background, two speakers (simultaneous) 1 male, 1 female, – 1 signal
- speech with noisy background (cocktail party), multiple speakers – 1 signal
- speech, single speaker, musical (instrumental) background, male – 1 signal
- singing voice, male, clean background – 2 signals
- singing voice, female, clean background – 2 signals
- singing voice, male, musical (instrumental) background – 2 signals

- singing voice, female, musical (instrumental) background – 2 signals
- singing voice, male, complex (instrumental + choir) background – 1 signal
- pure music (instrumental) – 3 signals

The signals were recorded initially at 44.1 kHz sample frequency, linear PCM, 16 bits/sample then, different replicas were resampled at 32, 16 and 8 kHz, obtaining 76 original signals. For the playback we used a PC which delivers uncompressed digital signal on SPDIF interface. The digital to analog conversion and subsequent amplification are performed using a Sony STR DB780 QS equipment. For speakers we used Yamaha NS10M Studio monitors. A group of 5 trained listeners for subjective quality evaluation was formed. Sounds were scored in pairs.

Because the proposed algorithm is based on the KBS-TSM solution, we compare the time-scaled replicas generated by these two algorithms. To quantify the improvements brought by each stage of the new algorithm, for the same original signal (O signal) more time-scaled replicas were produced. The first replica is delivered by the proposed algorithm (R signal) while the following ones (T1, T2, T3, T4 signals) are obtained using the same algorithm, but in which one of the blocks (at a time) was replaced with the corresponding original procedure from the KBS-TSM algorithm. More exactly, T1 was obtained using the KBS-TSM segment alignment and concatenation algorithm. For T2 signals, the "Frame Selection" block was replaced. In T3 experiments, an ordinary STFT was used instead of our IF analysis and finally, the T4 signals are generated by changing the IF attractors from our algorithm with a peak extraction routine as used in [5].

In each experiment we present to the listener the O signal, the R signal and one of the T signals. The experiment is blind (the listener knows which is the O signal but doesn't know which is the R or T signal). The listener can review freely the original and processed sounds and then scores the time-scaled signals. Scores were assigned on a comparative category rating scale of -3 to 3 (see Table 2)[17]. The tests are repeated for different TSM ratios. The results are presented in Table 2.

Table 2. Comparative MOS (CMOS) scale

Score	Comparative Quality
3	Proposed method much better than KBS-TSM algorithm
2	Proposed method better than KBS-TSM algorithm
1	Proposed method slightly better than KBS-TSM algorithm
0	Proposed method equal to KBS-TSM algorithm
1	Proposed method slightly worse than KBS-TSM algorithm
2	Proposed method worse than KBS-TSM algorithm
3	Proposed method much worse than KBS-TSM algorithm

The first modification (T1 signals) refers to correlative concatenation of segments. The experiments and listening tests show considerable improvements if the proposed algorithm is adopted. This is due to a more consistent approach in choosing the correlation frames and due to a more robust estimator for cross-correlation.

Table 3. CMOS results

TSM Ratio α	T1	T2	T3	T4
0.6	0.9	2.1	0.6	0
0.8	1.3	1.9	0.6	0
1.2	0.9	1.2	0.1	0
1.6	0.4	0.5	0	0

The listening tests on decoupling the synthesis frame length from the analysis windows length (T2) show a noticeable increase in the quality of the output signal. The solution offered by the KBS-TSM imposes equal lengths for the analysis and synthesis frames and thus an increase in the analysis resolution implies an increase in the synthesis duration. This causes reverberation artifacts to become evident.

The third modification (T3) refers to the use of IF spectrogram instead of ordinary STFT. The theoretical improvement is again verified by the experiments. By using a direct STFT, as in [5], the instantaneous frequencies of the partials are restricted to discrete values corresponding to the DFT/FFT bins. IF analysis allows fractional values for the normalized instantaneous frequencies and therefore enables a more accurate estimation of the amplitudes and phases of the partials.

By replacing the peak extraction routine with the IF attractors the number of partials is decreased (with approximately 10%). This allows a further decreasing of the arithmetic complexity in resynthesis stage. The

listening tests (T4) show that the complexity reduction doesn't alter the quality of the time-scaled signal. Despite the fact that for large values of TSM ratios over 1 the improvement becomes smaller, globally, the experimental results show the superiority of the proposed algorithm for all TSM ratios.

7. CONCLUSIONS

This paper follows the trend of merging aspects of time domain techniques with improved spectral analysis/synthesis methods used for sinusoidal modeling. We proposed a new algorithm which offers partial remedies for the deficiencies met at the KBS-TSM algorithm. This was possible through a careful construction of the analysis frames used in the correlative concatenation stage accompanied by a more suitable selection of the correlation estimator. The entire timbre morphing procedure was also revised. Using the IF spectrogram, we refined the spectral analysis and we performed a more accurate estimation of the limit conditions for the frequencies, the amplitudes and the phases of the partials required by the sinusoidal model. The smearing artifacts were kept under control by reducing the length of the synthesis frame, but we decoupled the synthesis and the analysis frames. This increased flexibility in the control of the location and the length of the frames, allowed us to properly balance the temporal and spectral resolution. The efficiency of the solutions included in the proposed algorithms was confirmed by the results of the performed listening comparative tests. The quality improvement offered by our algorithm is significant for those TSM ratio values where the KBS-TSM algorithm is known not to operate properly. In addition, we used the IF attractors as an alternative for the common peak extraction procedure. This adopted solution allows us to decrease the arithmetic complexity by reducing the number of the partials involved in SM synthesis. This benefit was obtained with no modification of the perceived quality for the time-scale signals.

To sum it all up, the proposed algorithm could be used to extend, with an acceptable implementation cost, the high quality operating range beyond the typical ±15%, as it is for classical time domain algorithms.

REFERENCES

1. KAHRS M., BRANDENBURG K., *Applications of digital signal processing to audio and acoustics*, Kluwer Academic Publishers, 1998.
2. ROUCOS S., WILGUS A., *High Quality Time-Scale Modification of Speech*, Proceedings of International Conference on Acoustics Speech and Signal Processing, pp.493-496, 1985.
3. DUDLEY H., *Remaking Speech*, Journal of Acoustical Society of America, **Vol. 11**, No.2, pp.169-177, 1939.
4. LAROCHE J., DOLSON M., *Improved Phase Vocoder Time-Scale Modification of Audio*, IEEE Transactions on Speech and Audio Processing, **Vol. 7**, No.3, pp.323-332, 1999.
5. KNOX D. BAILEY N. STEWART I., *A Simple Hybrid Approach to the Time-Scale Modification of Speech*, Journal of Audio Eng. Soc., **Vol. 53**, No.7/8, pp.612-618, 2005.
6. HARTMAN W.M., *Signals, Sound, and Sensations*, Springer Verlag, New York, 1998.
7. McAULAY R.J., QUATIERI T.F., *A Speech Analysis/Synthesis Based on a Sinusoidal Representation*, IEEE Transactions on Acoustics, Speech and Signal Processing, **Vol. 34**, No.4, pp.744-754, 1986.
8. GEORGE E.B., SMITH M.J.T., *A New Speech Coding Model Based on a Least-squares Sinusoidal Representation*, ICASSP-87 Proceedings, pp.1641-1644, 1987.
9. GEORGE E.B., SMITH M.J.T., *Analysis-by-synthesis / Overlap-add sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones*, Journal of Audio Eng. Soc., **Vol. 40**, No.6, pp.497-516, 1992.
10. MALLAT S.G., ZHANG Z., *Matching Pursuits With Time-frequency Dictionaries*, IEEE Transactions on Signal Processing, **Vol. 41**, No.12, pp.3397-3415, 1993.
11. ETERMOGLU C. CUPERMAN V., *Matching Pursuits Sinusoidal Speech Coding*, IEEE Transactions on Speech and Audio Processing, **Vol. 11**, No.5, pp.413-424, 2003.
12. DORRAN D., LAWLOR R., COYLE E., *A Hybrid Time-Frequency Approach to Audio Time-Scale Modification*, Journal of Audio Eng. Soc., **Vol. 54**, No.1/2, pp.21-31, 2006.
13. JANG H.K., PARK J.S., *Multiresolution Sinusoidal Model With Dynamic Segmentation for Timescale Modification of Polyphonic Audio Signals*, IEEE Transactions on Speech and Audio Processing, **Vol. 13**, No.2, pp.254-262, 2005.
14. ABE T., KOBAYASHI T., IMAI S., *Harmonic Tracking and Pitch Extraction Based on Instantaneous Frequency*, Proceeding of ICASSP 95, **Vol. I**., Detroit, pp. 756-759, 1995.
15. ABE T., KOBAYASHI T., IMAI S., *The IF Spectrogram: A New Spectral Representation*, Proceeding of International Symposium on Simulation Visualization and Auralization for Acoustic Research and Education, ASVA'97, Tokyo, Japan, pp.423-430, 1997.
16. ABE T., HONDA M., *Sinusoidal Model Based on Instantaneous Frequency Attractors*, IEEE Transactions on Audio, Speech and Language Processing, **Vol. 14**, No.4, pp.1292-1300, 2006.
17. ITU-T, *P800 – Methods for Subjective Determination of Transmission Quality*, ITU-T Telecommunication sector, 1996.