Time-Scale Modification of Audio and Speech Signals

Eero-Pekka Damskägg

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 22.1.2018

Thesis supervisor:

Prof. Vesa Välimäki

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Eero-Pekka Damskägg

Title: Time-Scale Modification of Audio and Speech Signals

Date: 22.1.2018 Language: English Number of pages: 7+56

Department of Signal Processing and Acoustics

Professorship: Audio Signal Processing

Supervisor and advisor: Prof. Vesa Välimäki

In audio time-scale modification (TSM), the duration of an audio recording is changed while retaining its local frequency content. In this thesis, a novel phase vocoder based technique for TSM was developed, which is based on the new concept of fuzzy classification of points in the time-frequency representation of an input signal. The points in the time-frequency representation are classified into three signal classes: tonalness, noisiness, and transientness. The information from the classification is used to preserve the distinct nature of these components during modification. The quality of the proposed method was evaluated by means of a listening test. The proposed method scored slightly higher than a state-of-the-art academic TSM technique, and similarly as a commercial TSM software. The proposed method is suitable for high-quality TSM of a wide variety of audio and speech signals.

Keywords: audio systems, digital signal processing, music, spectral analysis, spectrogram, speech processing

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Eero-Pekka Damskägg

Työn nimi: Audio- ja puhesignaalien aika-asteikon muuttaminen

Päivämäärä: 22.1.2018 Kieli: Englanti Sivumäärä: 7+56

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Audiosignaalinkäsittely

Työn valvoja ja ohjaaja: Prof. Vesa Välimäki

Äänen aika-asteikon muuttamisessa äänitteen pituutta muokataan niin, että sen paikallinen taajuussisältö säilyy samanlaisena. Tässä diplomityössä kehitettiin uusi, vaihevokooderiin pohjautuva menetelmä äänen aika-asteikon muuttamiseen. Menetelmä perustuu äänen aikataajuusesityksen pisteiden sumeaan luokitteluun. Pisteet luokitellaan soinnillisiksi, kohinaisiksi ja transienttisiksi määrittällä jatkuva totuusarvo pisteen kuulumiselle kuhunkin näistä luokista. Sumeasta luokittelusta saatua tietoa käytetään hyväksi näiden erilaisten signaalikomponenttien ominaisuuksien säilyttämiseen aika-asteikon muuttamisessa. Esitellyn menetelmän laatua arvioitiin kuuntelukokeen avulla. Esitelty menetelmä sai kokeessa hieman paremmat pisteet kuin viimeisintä tekniikkaa edustava akateeminen menetelmä, ja samanlaiset pisteet kuin kaupallinen ohjelmisto. Esitelty menetelmä soveltuu monenlaisien musiikki- ja puhesignaalien aika-asteikon muuttamiseen.

Avainsanat: audiojärjestelmät, digitaalinen signaalinkäsittely, musiikki, puheenkäsittely, spektrianalyysi, spektrogrammi

Preface

This Master's thesis work was carried out in the Department of Signal Processing and Acoustics at Aalto University with funding from the Aalto University School of Electrical Engineering. The funding period was from March 2017 throughout September 2017.

I would like to thank Professor Vesa Välimäki for his excellent guidance during this project, as well as Mr. Etienne Thuillier for sharing his expertise on this topic with me. I would also like to thank the experience director of the Finnish Science Center Heureka Mikko Myllykoski, who gave the idea for this thesis. Finally, I would like to thank all the people at Aalto Acoustics Lab for providing a pleasant and fun working environment.

Helsinki, 18.1.2018

Eero-Pekka Damskägg

Contents

Symbols and Abbreviations

Symbols

H_a	analysis hop size
H_s	synthesis hop size
k	DFT bin number
K	DFT size (number of bins)
m	frame time index
n	discrete time
N	frame size
R_n	noisiness
R_s	tonalness
R_t	transientness
r_t	frame transientness
$\mathbb{R}$	set of all real numbers
w_a	analysis window function
w_s	synthesis window function
x	input signal
X	analysis STFT
y	output signal
Y	synthesis STFT
$\mathbb{Z}$	set of all integers
κ	heterodyned phase increment
ω	normalized frequency
ω_{inst}	instantaneous frequency

Abbreviations

COLA	constant overlap-add
DFT	discrete Fourier transform
FBS	filter-bank summation
FFT	fast Fourier transform
OLA	overlap-add
PSOLA	pitch-synchronized overlap-add
STFT	short-time Fourier transform
SOLA	synchronized overlap-add
TSM	time-scale modification
WSOLA	waveform-similarity overlap-add

1 Introduction

Time-scale modification (TSM) is an audio signal processing technique which alters the duration of an audio signal while retaining its local frequency content [1, 2, 3]. TSM has many applications, such as fast browsing of speech recordings [4], music production and DJing [5, 6], foreign language learning [7], fitting a piece of music to a prescribed time slot [8], and slowing down the soundtrack for slow-motion video [9]. Additionally, TSM is often used as a processing step in pitch shifting, which aims at changing the local frequency content of the signal, while preserving its duration [2, 3, 8, 10, 11].

By itself, changing the duration of a sound by some factor is a straightforward task. Changing the duration in such a way, that the modified audio corresponds to the expectations of the listener however, is not. As an example, consider an orchestra playing a piece of music. When modifying the time scale of such a signal, it is typically desired for the modified audio to sound as if the orchestra was actually playing the piece faster or slower. Thus, it seems like high-level information about what different instruments sound like when played at different speeds is needed in order to carry out such a transformation.

For the bowed string instruments, such as the violin, analyzing the time-varying frequency content of the sound, and re-synthesizing a sound of different duration with the same frequency content seems like a reasonable approach for natural-sounding TSM. Consider the snare drum however. As opposed to the sustained notes played by the violin, playing of the snare drum consists of short staccato sounds. Playing a snare drum pattern with a different speed has little or no effect on how the individual drum hits sound like. Thus, TSM of such a sound should consist of separating the individual drum hits and moving them in time according to the desired modification factor. Finally, many musical sounds, such as the notes played on the piano or on the glockenspiel, consist of a sharp impulsive sound, followed by a sustained tonal part. When modifying the time scale of such sounds, the duration of the sharp attack part, that is, the transient, should be preserved, whereas the duration of the tonal part should be altered according to the desired modification factor.

Most TSM techniques fall into two main categories: time-domain techniques, and time-frequency-domain techniques. Standard time-domain techniques, such as the synchronized overlap-add (SOLA) [12], the waveform-similarity overlap-add (WSOLA) [13], and the pitch-synchronous overlap-add (PSOLA) [14] are most suitable for TSM of quasi-harmonic signals. That is, signals which can be considered as a sum of slowly-varying sinusoids with frequencies that are roughly harmonically related to each other. In the above discussion, the sustained notes played by the violin, and the tonal part of the notes played by the piano and the glockenspiel fall into this category. Considering transient sounds however, some modifications to the standard methods are needed in order to achieve a meaningful transformation. Furthermore, time-domain techniques suffer from artifacts when applied to polyphonic signals, since they are only able to preserve the most dominant periodicity in the input signal [3].

Time-frequency-domain TSM techniques are typically based on the phase vocoder

(e.g. [15, 16, 17]), which was originally introduced by Flanagan and Golden in 1966 [18]. By means of analyzing the short-time spectra of the sound, phase vocoder based techniques are able to preserve all the periodicities in the input signal. Therefore, these techniques can provide a meaningful transformation of the time-scale even for polyphonic signals. However, since the processing in the phase vocoder is based on a sinusoidal model of the input signal, it is most suitable for modifying sounds which can be considered as a sum of slowly-varying sinusoids. Transients processed with the phase vocoder suffer from a softening of the perceived attack, often referred to as "transient smearing" [2, 3, 19]. Thus, similarly as with time-domain methods, transients need to be handled separately in order to increase the subjective quality of the transformation.

In this thesis, existing literature on TSM techniques is reviewed, and a novel TSM technique is developed which addresses some of the problems arising in these techniques. The proposed TSM technique relies on the new concept of fuzzy classification of points in the time-frequency representation of the input signal. The points are assigned to three signal classes: tonalness, noisiness, and transientness. Each time-frequency point belongs to all of the classes simultaneously, with a certain degree of membership for each class. The information from the classification is used to preserve the subjective quality of these distinct signal classes during TSM. To evaluate the quality of the proposed method, a listening test was conducted. The results of the listening test suggest that the proposed method is competitive against a state-of-the-art academic TSM method and a commercial TSM software.

The remainder of this thesis is structured as follows. In Section 2, a few of the most notable time-domain techniques for TSM are reviewed and discussed. In Section 3, fundamentals of phase-vocoder-based TSM are presented, and typical problems arising in such processing are introduced. In Section 4, extensions to the standard phase-vocoder-based processing which alleviate some of the typical problems, are reviewed. In Section 5, the developed TSM technique is presented. In Section 6, performance of the proposed technique is evaluated by means of a listening test. Finally, Section 7 concludes the thesis.

2 Time-Domain Techniques for Time-Scale Modification

Time-domain techniques for TSM are based on an analysis and synthesis procedure. In the analysis, the input audio is split into short segments. Each segment contains the local frequency content of the signal around the analysis time instant. In the synthesis, the frames are relocated in time, such that either a time contracted or a time expanded signal is obtained. The time-segments are also often modified by means of applying a weighting with a window function. There are various techniques which use this approach, each with slight differences on how these steps are carried out. All of these techniques fall under the overlap-add (OLA) family.

2.1 Standard Overlap-Add

The standard OLA procedure is visualized in Figure 1. As shown in Figure 1a, the procedure begins by selecting a time-limited analysis frame $x_m[n]$ from the input signal $x[n]$. Next, an analysis window $w[n]$ is applied to the analysis frame to obtain the synthesis frame $y_m[n]$. Then, the synthesis frame is added to the output signal $y[n]$ (Figure 1b). The next analysis frame $x_{m+1}[n]$ is selected from the input signal at a specific distance H_a, which is known as the analysis hop size (Figure 1c). The analysis window is applied to the new analysis frame to obtain the next synthesis frame $y_{m+1}[n]$. Finally, the synthesis frame is added to the output signal at a distance from the previous synthesis frame determined by the synthesis hop size H_s (Figure 1d). The amount of time expansion or contraction is denoted by the TSM factor α, which is determined by the ratio between the analysis and synthesis hop sizes:

$$\alpha = \frac{H_s}{H_a}, \tag{1}$$

such that $H_s > H_a$ results in time expansion, while $H_s < H_a$ results in time contraction of the input signal.

A digital implementation of the OLA technique for TSM was first introduced by Lee [20]. The implementation is based on a ring buffer. A write pointer moves around the buffer at some speed, writing values of the input signal to the buffer. Simultaneously, a read pointer moves around the buffer at a speed equal to the systems output sampling rate, reading values from the memory to the output. A time-contracted output signal is obtained if the write pointer moves faster than the read pointer, and a time-expanded signal is obtained if the write pointer moves slower than the read pointer. In this approach, the transition to a new segment in the output happens when the read and write pointers cross each other.

This simple method suffers from three artifacts: discontinuities, pitch distortions, and transient duplication or skipping. Discontinuities occur at the transitions between two segments. Two consecutive segments are unlikely to be aligned in a way that there is a smooth continuation from the end of one segment to the start of the next one. This problem is avoided by choosing an analysis window which tapers to zero at the edges of the analysis frame. However, pitch distortions still occur when the

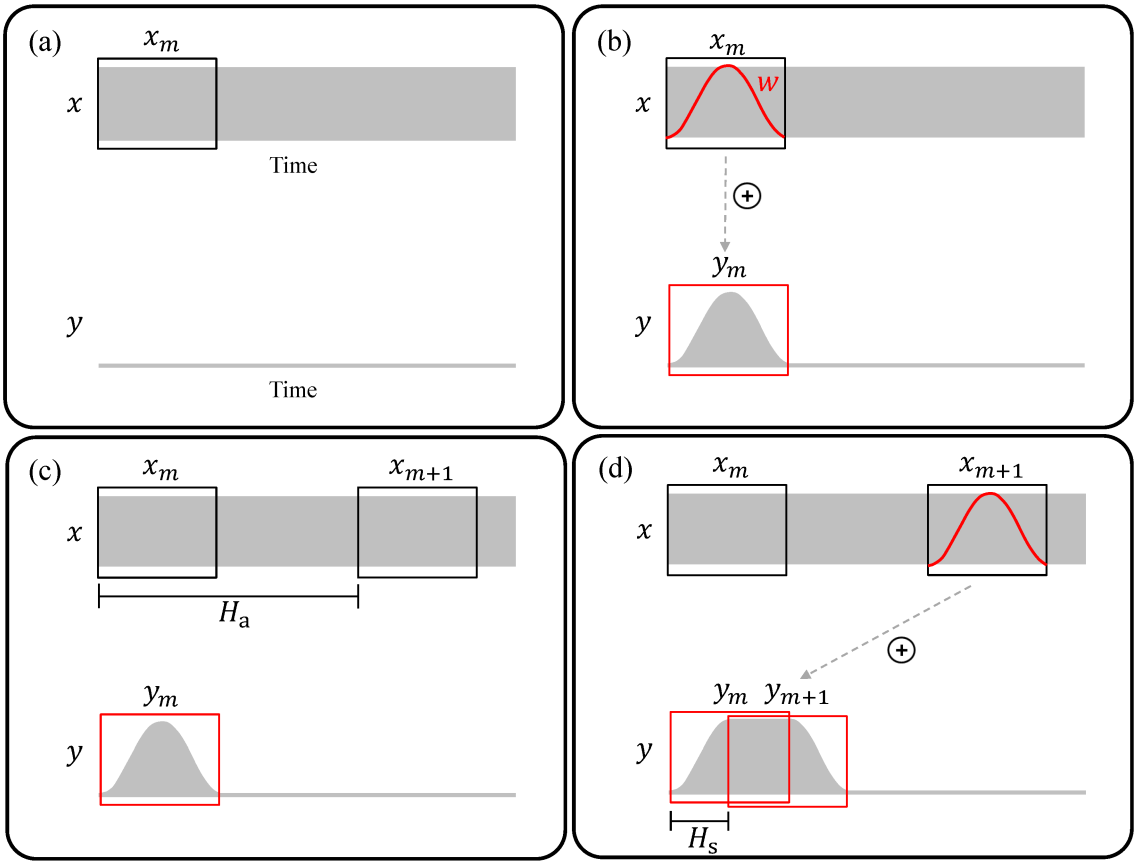


Figure 1: The OLA procedure for TSM. (**a**) An analysis frame $x_m[n]$ is segmented from the input signal. (**b**) The analysis frame is weighted by an analysis window $w[n]$ to obtain the synthesis frame $y_m[n]$, which is added to the output signal. (**c**) The next analysis frame $x_{m+1}[n]$ is picked at a distance H_a, denoted as the analysis hop size, from the previous analysis time instant. (**d**) The new analysis frame is weighted by the analysis window, and added to the output signal, at a distance H_s, denoted as the synthesis hop size, from the previous synthesis time instant. This example is of time contraction, as the synthesis hop size is smaller than the analysis hop size. The figure has been adopted from [3].

pitch periods in two consecutive segments are not aligned. That is, when a transition happens from one segment to the next, the output waveform suddenly jumps from one position in the pitch period to another. This can alter the perceived pitch in the output signal. Transient duplication occurs in time expansion when multiple time-shifted synthesis frames contain the same transient due to the analysis hop size being shorter than the synthesis hop size. Conversely, transient skipping can occur in time contraction, when the analysis hop size is greater than a short transient event, and the frames skip the transient location during analysis.

2.2 Overlap-Add Variants

To overcome the limitations of the OLA technique, several improvements to the standard technique have been proposed. In the following, a few of the most notable ones are reviewed.

2.2.1 Synchronous Overlap-Add

In order to solve the issues with the standard OLA procedure for TSM, some sensitivity to the input signal needs to be introduced to the processing. To smoothly transition from one synthesis frame to the next, the pitch periods in the overlapping regions of two consecutive frames should be aligned. In the synchronous overlap-add (SOLA) algorithm [12], this is achieved by allowing some flexibility in choosing the time region in which the transition from one frame to the next occurs. The SOLA technique is visualized in Figure 2. First, similarly as in the standard OLA procedure, the input signal is split into partially overlapping analysis frames $x_m[n]$, using a fixed analysis hop size H_a and an analysis frame length N, as shown in Figure 2a. No fixed analysis window is applied to the analysis frames to obtain the synthesis frames. The synthesis is initialized by adding the first analysis frame to the beginning of the output signal:

$$y[n] = x_0[n], \text{ for } n = 0, 1, ..., N-1, \tag{2}$$

where N is the analysis frame length.

The following analysis frames $x_m[n]$ are synchronized to the neighborhood $y[n + mH_s]$ of the output signal on a frame-by-frame basis. That is, the new frame is first moved to a fixed location determined by the frame index m and the synthesis hop size H_s, as shown in Figure 2b. Next, the frame is time-shifted slightly, such that the preceding synthesized output signal and the added frame are maximally similar in their overlapping region, as shown in Figure 2c. The optimal location for adding the new frame is found as time lag Δ, which gives the maximum of the normalized cross-correlation between the two signals:

$$r_m[\Delta] = \frac{\sum_{n=0}^{L-1} y[mH_s + \Delta + n] x_m[n]}{\left(\sum_{n=0}^{L-1} y^2[mH_s + \Delta + n] \sum_{n=0}^{L-1} x_m^2[n] \right)^{1/2}}, \tag{3}$$

where L is the length of the overlap between the time-shifted analysis frame and the synthesized output signal. Denoting the maximum of the cross-correlation by Δ_m, the new frame is then added to the output as follows:

$$y[n + mH_s + \Delta_m] = \begin{cases} (1 - f[n]) y[n + mH_s + \Delta_m] + f[n] x_m[n], & \text{for } 0 \leq n \leq L_m - 1 \\ x_m[n], & \text{for } L_m \leq n \leq N - 1, \end{cases} \tag{4}$$

where L_m is the length of overlap between the synthesized signal and the analysis frame further time-shifted by Δ_m. Here, $f[n]$ is a weighting function obtaining values in the interval $[0, 1]$ which is used to smoothly transition to the new frame in the

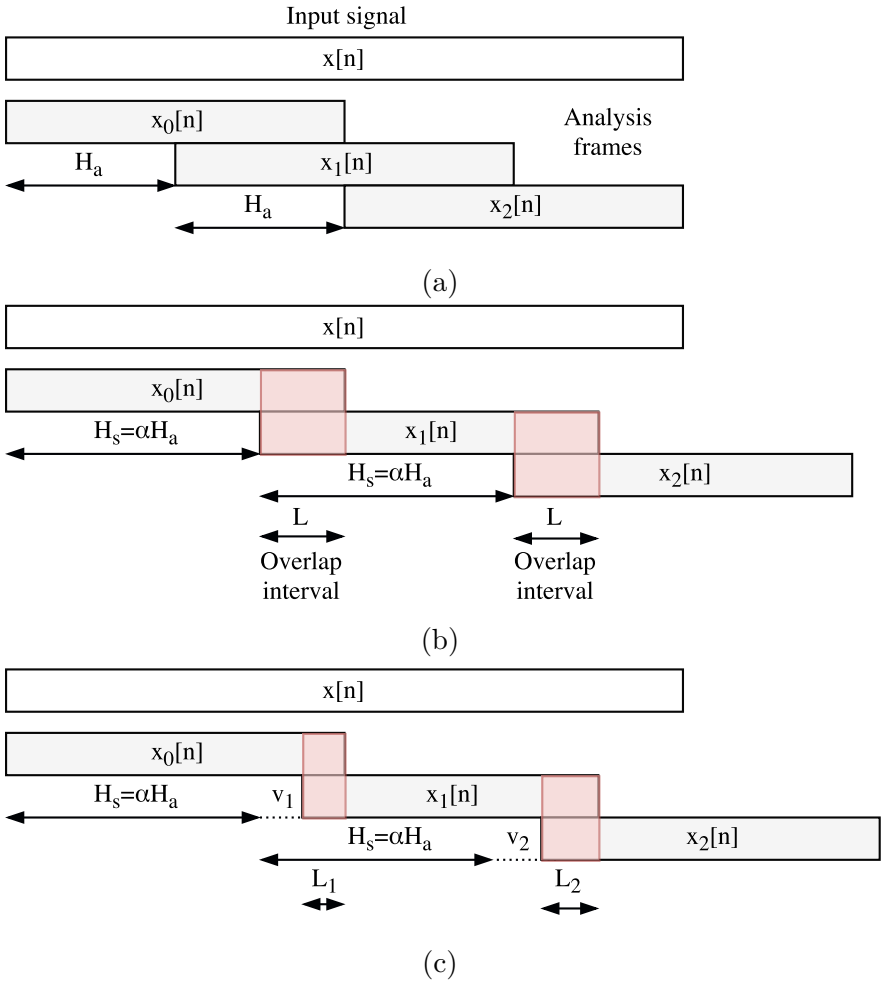


Figure 2: Synchronized overlap-add (SOLA) time-scale modification (TSM). (**a**) The input signal is segmented into analysis frames, which are computed at a rate determined by the analysis hop size H_a. (**b**) The frames are relocated in time according to the TSM-factor α, leaving an overlap of length L between two consecutive frames. (**c**) Inside the overlap interval of two consecutive frames, the position of the latter frame is adjusted such that the periodic structures in the frames are maximally aligned. The figure has been adapted from [8].

output. A standard choice for the weighting, given an overlap of length L_m, is the linear function

$$f[n] = \frac{n}{L_m - 1}, \text{ for } 0 \leq n \leq L_m - 1. \tag{5}$$

Proper alignment of pitch periods in consecutive frames in the synthesized output signal is based on the search of the maximum cross-correlation in Equation (3). The maximum cross-correlation is likely to occur when the pitch periods of the added frame, and the existing synthesized signal are aligned. Naturally, the approach only allows preserving the most dominant periodicity in the input signal, as that will have the largest effect on the cross-correlation function. Thus, the SOLA technique is not applicable to TSM of polyphonic signals, as pitch distortions will still remain on the

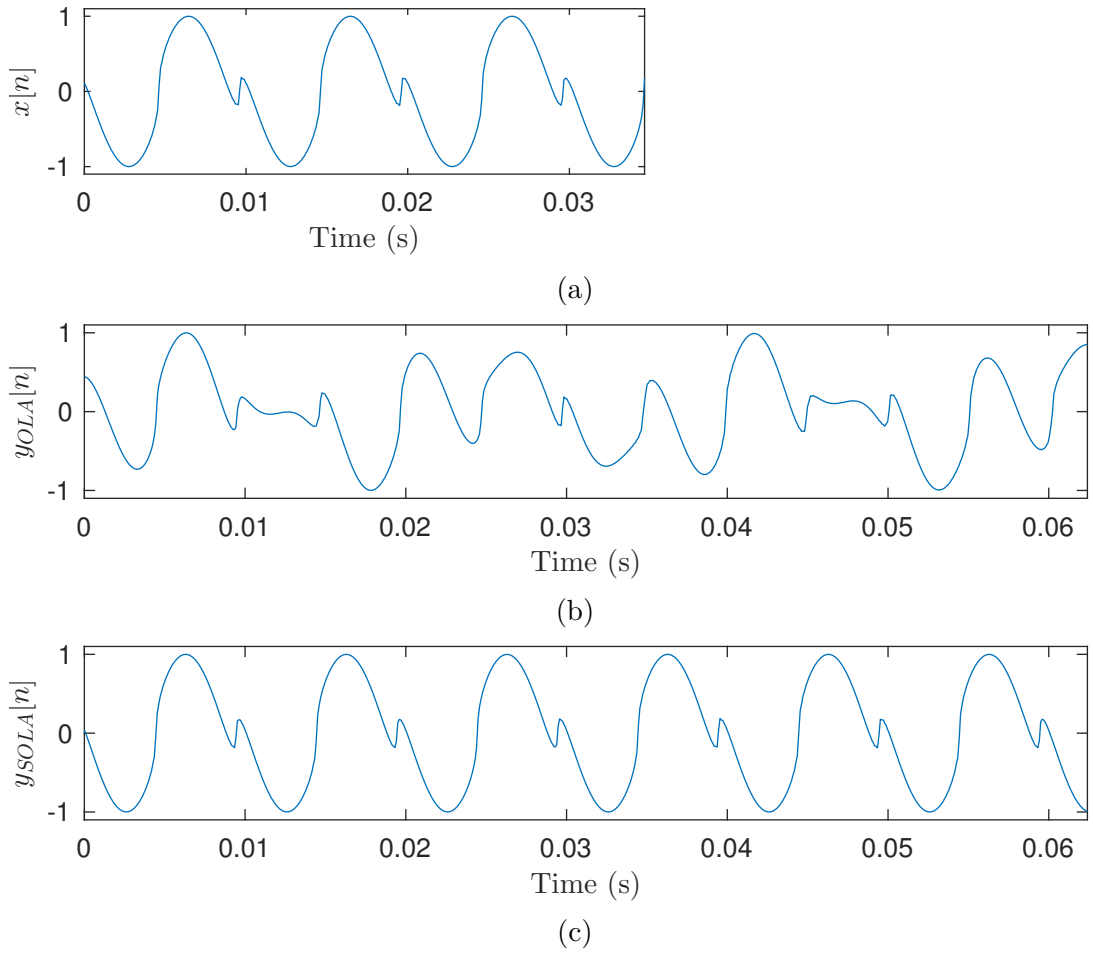


Figure 3: (**a**) A harmonic input signal and its time-scale modified versions with factor $\alpha = 1.8$, using (**b**) standard OLA, and (**c**) SOLA. The standard OLA technique is unable to preserve the shape and periodicity of the original waveform, whereas the SOLA technique preserves both.

periodic components besides the most dominant one.

Figure 3 shows the results of TSM with the OLA and SOLA techniques when applied to the harmonic signal shown in Figure 3a. A TSM-factor $\alpha = 1.8$ was used for the example. As seen in Figure 3b, the OLA technique is unable to preserve the waveform shape and fundamental frequency of the input signal. Since the technique works on fixed analysis and synthesis hop sizes, it has no sensitivity to the periodicities in the input signal. Thus, the harmonic waveforms in consecutive synthesis frames are in arbitrary phase to each other, and either constructive or destructive interference occurs at their overlapping segment. Conversely, in the SOLA technique, the synthesis hop size is adjusted on a frame-by-frame basis, such that each new synthesis frame added to the output signal is maximally aligned with the preceding waveform. Therefore, the SOLA technique is able to preserve the waveform shape and the fundamental frequency of the harmonic input signal, as can be seen in Figure 3c.

The SOLA technique is most applicable to signals, which contain a clear periodicity, such as the one in the example of Figure 3. However, most natural sounds, such as music and speech, also contain transients which play an important role in the perceived sound. To improve the quality of SOLA TSM with signals containing transients, techniques with a time-varying TSM factor have been suggested. In [21, 22], the transient and steady portions of the input signal are separated, and the TSM is only applied to the steady portions, while the transient portions are left unaffected. Because the original duration of the transient portions is preserved, the time scale in the steady portions needs to be contracted or expanded excessively in order to maintain the desired TSM factor. These methods were shown to increase the intelligibility of modified speech when compared to the standard SOLA procedure.

2.2.2 Pitch-Synchronous Overlap-Add

The pitch-synchronous overlap-add (PSOLA) [14, 23, 24] is a modification to the standard OLA procedure which is widely used in voice and speech processing. Contrary to the fixed analysis rate of the OLA and SOLA procedures, in PSOLA, the segmenting of the audio signal to obtain the analysis frames is done in a pitch-synchronous manner. The analysis stage in the PSOLA technique is visualized in Figure 4. It consists of finding the pitch marks, which are the time instants where the input signal obtains its maximum value of each pitch period. The analysis frames are then obtained by multiplying the signal with a series of analysis windows time shifted to the pitch marks:

$$x_m[n] = w_m[n - t_m]x[n], \tag{6}$$

where $w_m[n]$ is the analysis window used for computing the mth analysis frame and t_m are the locations of the analysis pitch marks. The length of the analysis window is proportional to the local pitch period, typically such that the window length for the mth analysis frame is

$$N_m = \mu P_m, \quad \mu \in \{2, 4\}, \tag{7}$$

where P_m is the local pitch period. The cases $\mu = 2$ and $\mu = 4$ correspond to 50% and 75% overlap between successive analysis windows, respectively. During portions of the input signal where there is no clear periodicity, such as unvoiced portions of speech, a constant rate analysis is used.

TSM with the PSOLA technique is based on determining the locations of the synthesis pitch marks based on the TSM factor α and the locations of the analysis pitch marks. The time difference between the synthesis pitch marks at time n is chosen as the time difference between the analysis pitch marks at time n/α. After determining the locations of the pitch marks t_q, where q is the synthesis frame index, appropriate synthesis frames $y_q[n]$ need to be computed for each pitch mark. This is illustrated in Figure 5, where a mapping between the synthesis pitch marks and the corresponding analysis time instants is shown. Given a synthesis pitch mark location t_q, a straightforward solution is to directly use the analysis frame closest to the time t_q/α. An alternative solution is to take the weighted average of the two

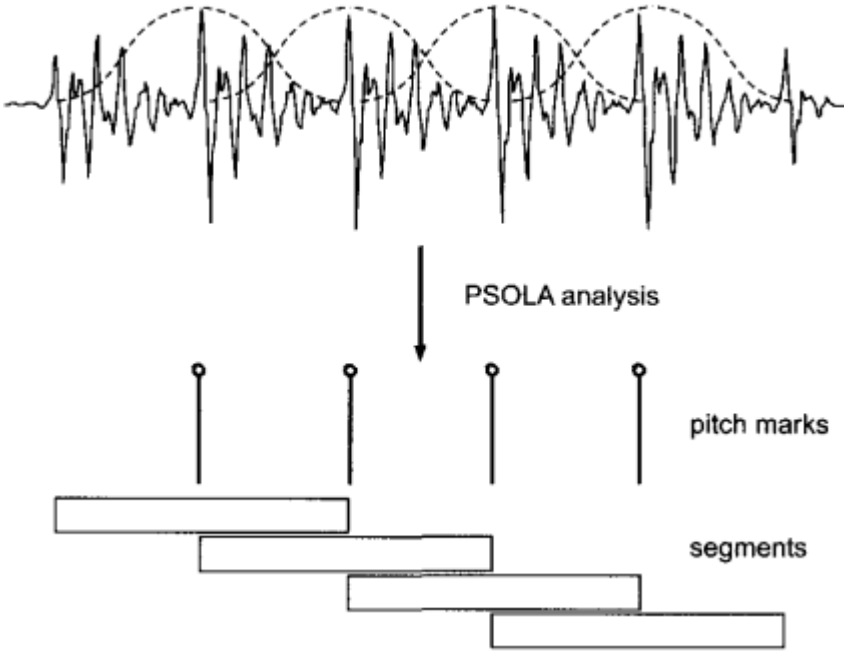


Figure 4: PSOLA analysis. First, pitch marks are extracted from the input signal. Next, the analysis frames are obtained by multiplying the signal with a series of analysis windows time shifted to the pitch marks. In the figure, the window length is $\mu = 2$ times the local period P_m, which corresponds to 50% overlap between successive analysis frames. The figure has been adopted from [8].

closest analysis frames, according to their distance to the time t_q/α. Finally, the output signal can be obtained by the least-squares overlap-add synthesis scheme of Griffin and Lim [25]:

$$y[n] = \frac{\sum_{q\in\mathbb{Z}} y_q[n - t_q]w_q[n - t_q]}{\sum_{q\in\mathbb{Z}} w_q^2[n - t_q]}, \tag{8}$$

where $w_q[n]$ is the synthesis window at frame index q. The denominator compensates for the energy modifications over time that are caused by the time-varying window and hop sizes.

2.2.3 Waveform-Similarity Overlap-Add

An alternative means of adapting the OLA technique to the main periodicity in the input signal was proposed by Verhelst and Roelands [13]. Contrary to the SOLA procedure, where some flexibility was allowed in the position of the synthesis frames, the WSOLA is based on allowing some flexibility in the position of the analysis frames. In the standard OLA, the fixed time position of the mth analysis frame is given by mH_a. In the WSOLA, the fixed analysis time can be shifted by $\Delta_m \in [-\Delta_{max}, \Delta_{max}]$, where Δ_{max} is the absolute value of the maximum allowed time shift. The adapted analysis time of the mth frame is then given by $mH_a + \Delta_m$.

Figure 6 visualizes the principle of WSOLA. The adapted analysis frames

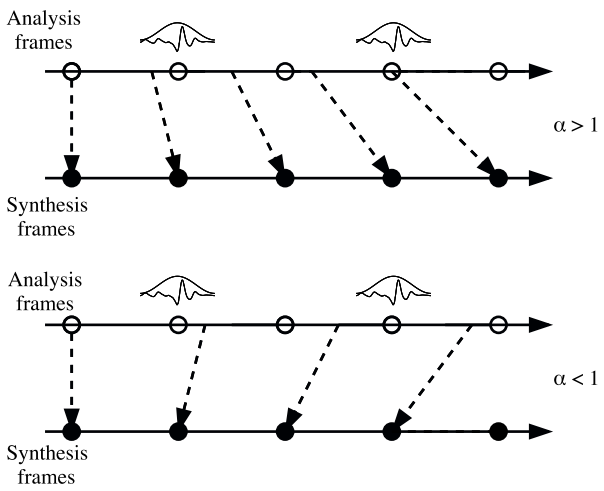


Figure 5: TSM with the PSOLA technique. The upper figure corresponds to time expansion, and the lower to time contraction. The analysis pitch marks t_m are shown as the unfilled circles, and the synthesis pitch marks t_q are shown as filled circles. A dashed line points from the corresponding analysis time to the synthesis pitch marks. The synthesis frame at that time instant is chosen either as the analysis frame closest to the corresponding analysis time instant, or as a weighted average of the two closest analysis frames. The figure has been adapted from [14].

windowed at the time shifted analysis positions are denoted by ${x'}_m$:

$$ {x'}_m[n] = \begin{cases} x[n + mH_a + \Delta_m], & \text{for } n \in [-N/2, N/2 - 1], \\ 0, & \text{otherwise}, \end{cases} \tag{9} $$

where N is the length of the analysis frames. The synthesis is initialized by adding the first adapted analysis frame to the beginning of the output signal:

$$ y[n] = {x'}_0[n], \text{ for } n \in [-N/2, N/2 - 1]. \tag{10} $$

Next, the problem is to find the optimal modified analysis time instant, such that the periodic structures of the new analysis frame ${x'}_{m+1}$ align optimally with the previously synthesized frame $y_m = {x'}_m$, when it is added at a distance corresponding to the synthesis hop size H_s from the previously synthesized frame. As shown in Figure 6b, if no constraints are applied on the possible analysis time instants, the optimal choice is at time $(m + 1)H_s$, which corresponds to the natural progression $\tilde{x}_m$ of the adjusted analysis frame ${x'}_m$. However, only variations up to Δ_{max} from the fixed analysis time instant are allowed when searching for the optimal frame position. Thus, the next adjusted analysis frame must be inside the extended frame region ${x^+}_m + 1$, which is visualized with the solid blue box in Figure 6b.

Similarly to SOLA, the optimal analysis time is found as the time shift which maximizes the normalized cross-correlation

$$ r_m[\Delta] = \frac{\sum_{n=-N/2}^{N/2-1} \tilde{x}_m[n] {x^+}_{m+1}[n + \Delta]}{\left(\sum_{n=-N/2}^{N/2-1} \tilde{x}_m[n]^2 \sum_{n=-N/2}^{N/2-1} {x^+}_{m+1}[n + \Delta]^2 \right)^{1/2}}. \tag{11} $$

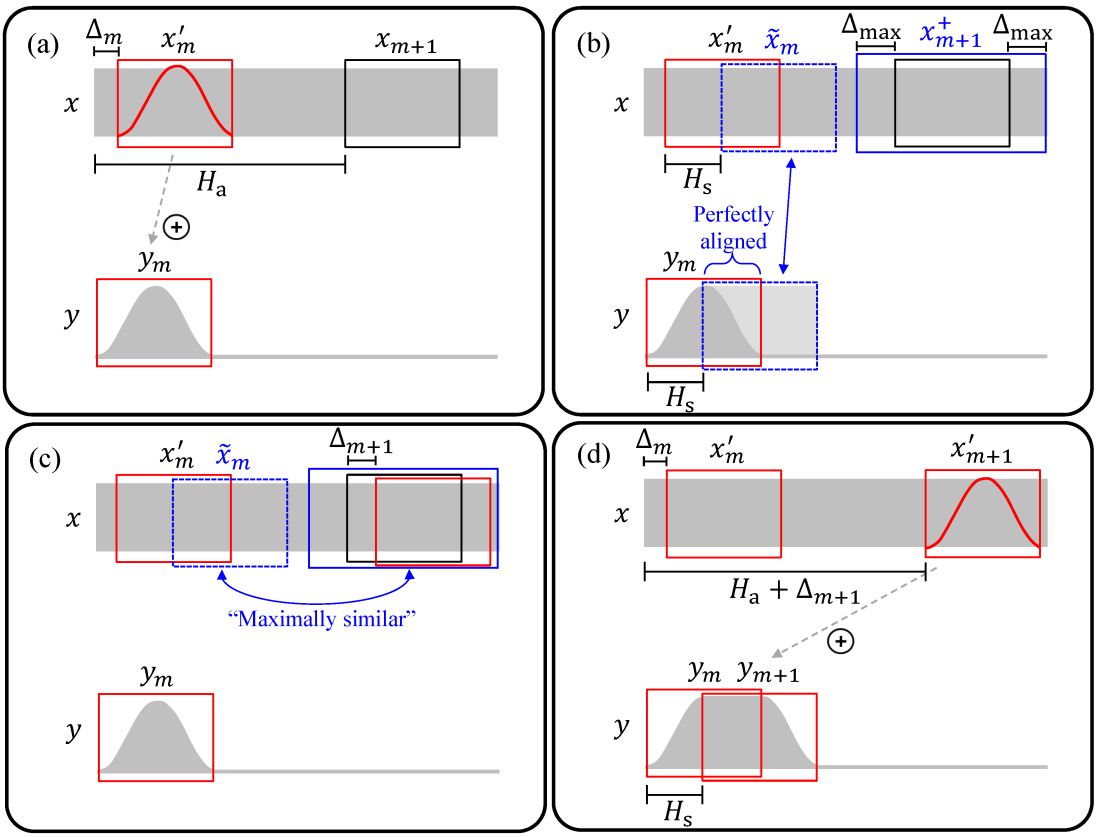


Figure 6: TSM with the WSOLA technique. (**a**) The synthesis is initialized by adding the first adjusted analysis frame $x'_m[n]$ to the beginning of the output signal. (**b**) The natural progression $\tilde{x}_m[n]$ is the analysis frame which is perfectly aligned with the previous synthesis frame. (**c**) The next analysis time position is searched for in the extended frame region $x^+_m + 1$, such that the computed analysis frame is maximally similar to the natural progression $\tilde{x}_m$ of the previous frame. (**d**) The new analysis frame is added to the output at a distance defined by the synthesis hop size H_s from the previous synthesis frame. The figure has been adopted from [3].

That is, the optimal adjusted analysis frame is considered as the frame inside the extended frame region $x^+_m + 1$, which is maximally similar to the natural progression $\tilde{x}_m$ of the adjusted analysis frame x'_m, as measured by cross-correlation (Figure 6c). Finally, as shown in Figure 6d, the output is the sum of the adjusted analysis frames synthesized at a fixed rate determined by the synthesis hop size.

WSOLA suffers from the some of the typical problems of the time-domain techniques. The transient duplication artifact in the case of time expansion with the WSOLA is visualized in Figure 7. As shown in the upper panel of Figure 7, during the analysis, a single transient is included in two consecutive analysis frames $x'_m[n]$ and $x'_{m+1}[n]$. During the synthesis, the frames are time shifted further apart from each other, which results in the transient being synthesized in two different time instants of the output signal. Furthermore, similarly to the time-domain techniques SOLA and PSOLA, the WSOLA is only able to preserve the most dominant periodicity in

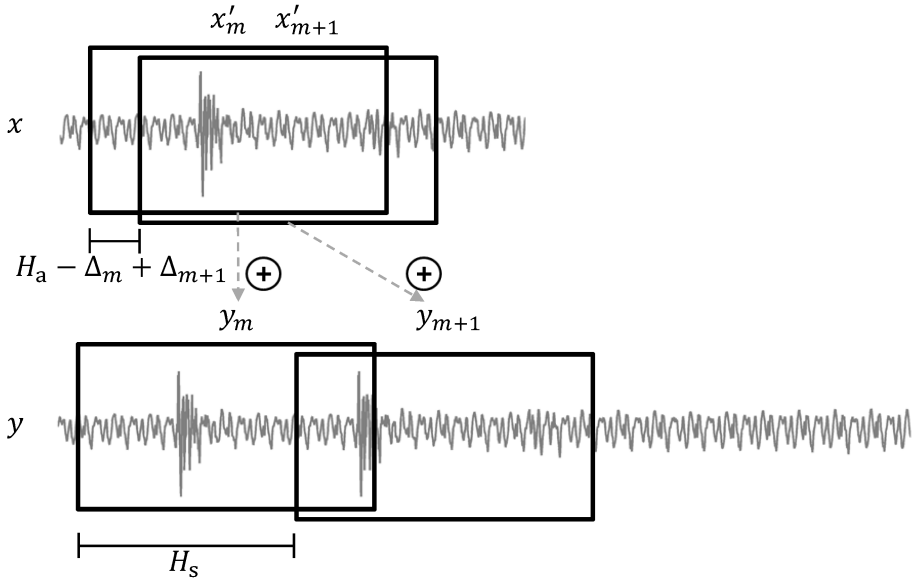


Figure 7: Transient duplication in time expansion with the waveform-similarity overlap-add (WSOLA) technique. The figure has been adopted from [3].

the input signal. Contrary to PSOLA, however, it does not require explicit pitch detection in order to preserve the periodicity in the synthesized signal. However, the analysis window size has to be selected such that at least a full period of the periodic pattern is captured in each frame.

3 Phase Vocoder Based Time-Scale Modification

As discussed in Section 2, time-domain methods, such as SOLA, WSOLA and PSOLA are only able to maintain the most dominant periodicity in the input signal during TSM. In order to improve the quality of TSM for signals which consist of multiple periodic components, the modification must be done in a way which preserves the periodicities of all signal components. This can be achieved with time-frequency processing. In time-frequency processing, the one-dimensional input signal is transformed into a two-dimensional representation of itself. Next, the obtained time-frequency representation is modified in some way. Finally the output signal is synthesized from the modified representation. This procedure is illustrated in Figure 8.

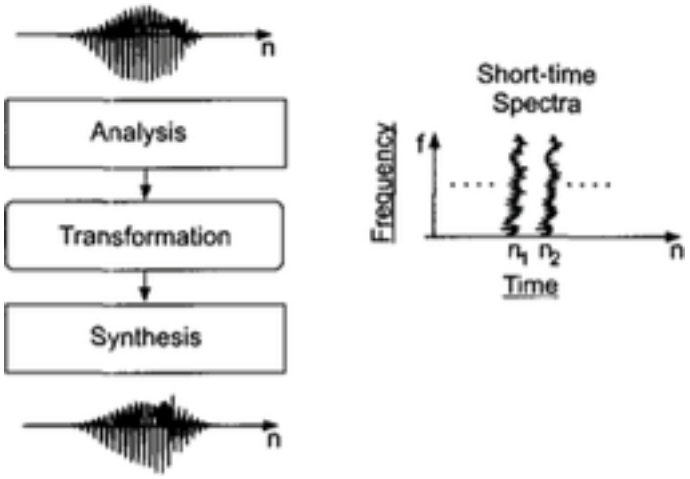


Figure 8: Time-frequency processing pipeline. The figure has been adopted from [26].

For audio and speech manipulation, time-frequency processing is typically based on the phase vocoder. In contrast to the magnitude-only representation of the original channel vocoder [27], the phase vocoder is based on the use of both magnitude and phase spectra. Thus, it allows perfect reconstruction of the original time-domain signal from the phase vocoder representation. The term phase vocoder was first introduced by Flanagan and Golden [18]. It refers to an analysis and synthesis procedure on a signal by means of its short-time Fourier transform (STFT). Additionally, in the phase vocoder, the rough frequency estimates of the STFT are made more accurate by combining information of two or more consecutive frames.

In 1976, Portnoff [28] introduced an implementation of the phase vocoder utilizing the fast Fourier transform (FFT) [29]. This significantly reduced the computational complexity involved, and allowed the phase vocoder to become a widely used tool for audio and speech processing. Another important improvement to the phase vocoder was the introduction of the phase difference method for computation of the instantaneous frequency [30], instead of the phase derivative method originally employed by Flanagan and Golden [18].

TSM using the phase vocoder is based on an analysis, modification and synthesis procedure, as illustrated in Figure 8. The analysis consists of obtaining a time-frequency representation of the signal by means of the STFT. The STFT representation is then modified to obtain the synthesis STFT. Finally, the output signal is obtained by synthesizing a time-domain signal from the modified representation by means of the inverse-STFT.

3.1 Short-Time Fourier Transform Analysis and Synthesis

3.1.1 Analysis

During the analysis, short-time spectra of the input signal are computed at a rate defined by the analysis hop size H_a, resulting in the analysis STFT [28, 31, 32, 33]:

$$X[m,k] = \sum_{n=-\infty}^{\infty} x[n]w_a[n - mH_a]e^{-j\omega_k n} \tag{12}$$

where $x[n]$ is the input signal, $w_a[n]$ is the analysis window function, and ω_k is the normalized center frequency of the kth vocoder channel, often also denoted as the kth bin. If the number of samples in the computed discrete Fourier transform (DFT) is given by K, the normalized center frequency of the kth bin is given by $\omega_k = 2\pi k/K$. The analysis window $w_a(n)$ is typically non-zero only in a finite interval $n = 0, ..., N-1$, with analysis window length N. When the analysis window length $N \leq K$, the sampling in the DFT will not cause time aliasing [34]. The mth analysis frame is centered around time $n = mH_a$. Given that the analysis window is non-zero only in a finite interval, practical computation of the STFT can be done by:

$$X[m,k] = \sum_{n=-N/2}^{N/2-1} x[n + mH_a]w[n]e^{-j\omega_k n}, \tag{13}$$

where the summation is done over a finite number of samples N. In this definition, the input signal is time shifted while the analysis window remains fixed around time 0.

The analysis STFT provides a time-frequency representation of the input signal. That is, the one-dimensional input signal $x[n]$ is represented by the complex-valued points $X[m,k]$ in the analysis STFT. For signal analysis and modification, the complex-valued STFT is usually decomposed to the magnitude and phase components, $|X[m,k]|$ and $\angle X[m,k]$, respectively, such that the original STFT can be retrieved by:

$$X[m,k] = |X[m,k]|e^{j\angle X[m,k])}, \tag{14}$$

where $|\cdot|$ gives the magnitude of a complex number and $\angle\cdot$ gives the angle.

The STFT can be interpreted using two different points of view: the OLA and filter-bank summation (FBS) points of view [34, 35]. One view is the Fourier dual of the other. Figure 9 illustrates the OLA interpretation of the STFT analysis. As

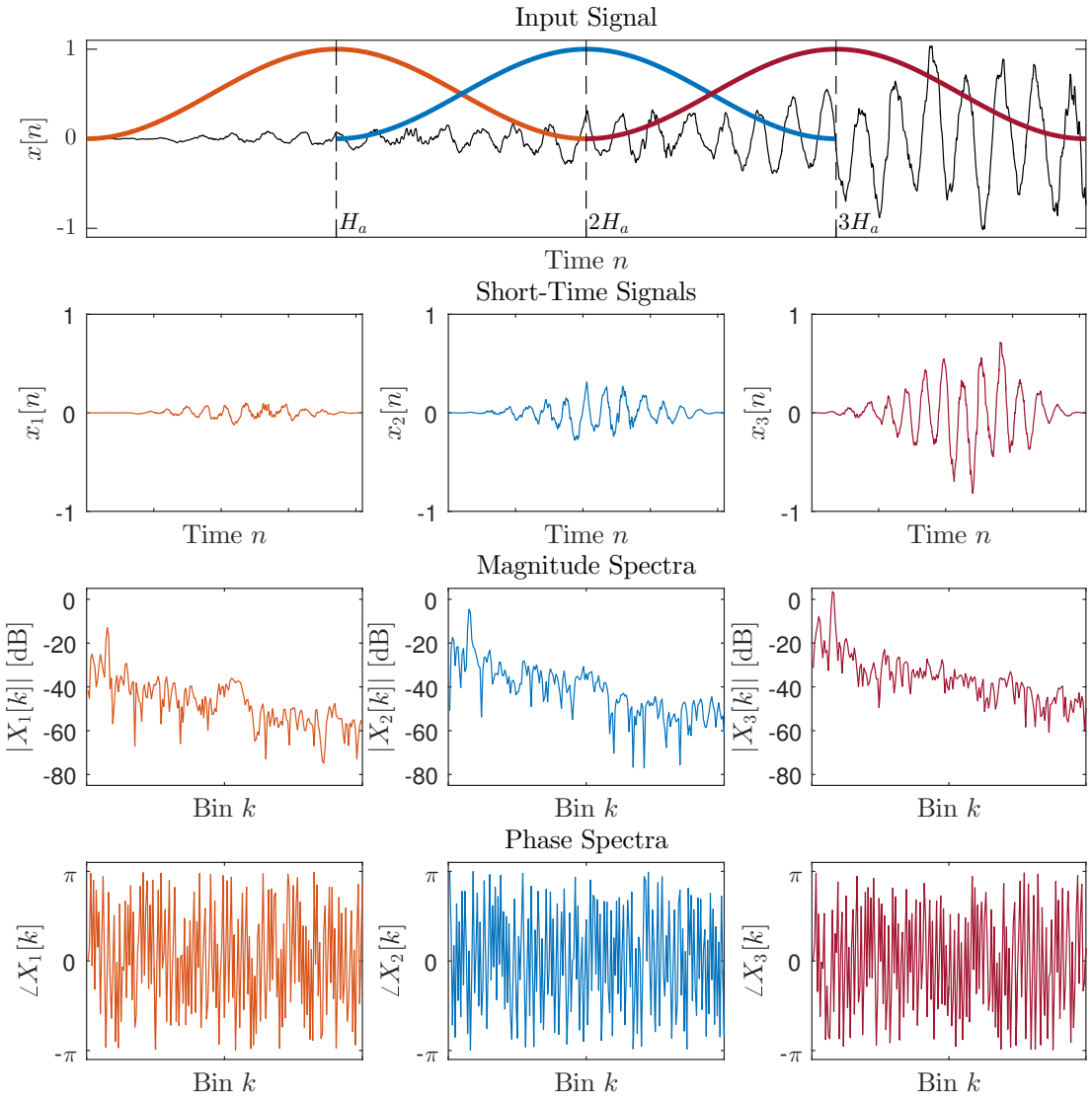


Figure 9: The OLA view of the STFT analysis.

shown in the top row of Figure 9, short-time signals are obtained from the input signal by a sample-wise multiplication with the time-shifted analysis window function:

$$x_m[n] = x[n]w_a[n - mH_a]. \tag{15}$$

In the figure, the analysis window is time shifted instead of the input signal for easier visualization. However, in practice the short-time signals are computed by:

$$x_m[n] = x[n + mH_a]w_a[n]. \tag{16}$$

The second row of Figure 9 shows the short-time signals obtained from the input signal in the top row. The short-time signals are computed at a rate determined

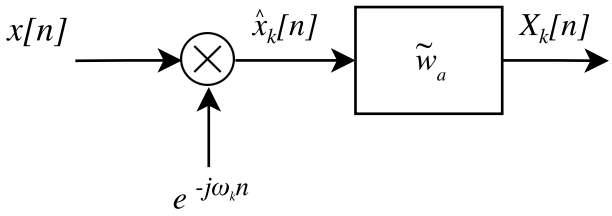


Figure 10: A single channel k of the STFT filter bank.

by the analysis hop size H_a. That is, a short-time signal is computed every H_ath sample. Next, the DFT is taken of each short-time signal:

$$X_m[k] = \sum_{n=-N/2}^{N/2-1} x_m[n] e^{-j\omega_k n}, \tag{17}$$

where it is assumed that the short-time signals have been computed as in Equation (16). This results in an interpretation of the STFT as a time sequence of short-time spectra $X_m[k]$. In the lower two rows in Figure 9, the computed magnitude and phase spectra of the short-time signals are shown.

An alternative way to interpret the the STFT is the FBS point of view. While in the OLA interpretation, the STFT is considered a time sequence of short-time spectra, in the FBS interpretation, the STFT is considered a frequency-ordered collection of narrow-band time-domain signals [34]. That is, the computation of the STFT in Equation (12) is interpreted as a parallel bank of K bandpass filters. In the computation of the STFT, the input signal is first modulated by the complex exponential in Equation (12), which results in the "heterodyned signal":

$$\hat{x}_k[n] = x[n] e^{-j\omega_k n}. \tag{18}$$

Modulation of the time-domain input signal with a complex exponential of frequency $-\omega_k$, is the Fourier dual of rotating the spectrum of the signal along the unit circle of the z plane towards zero frequency by ω_k. The heterodyned signals are then convolved with the time-reversed analysis window function $\tilde{w}_a$, which results in the output signal:

$$X_k[n] = (\hat{x}_k * \tilde{w}_a)[n], \tag{19}$$

where $\tilde{w}_a[n] = w_a[-n]$. The convolution by the time-reversed analysis window function corresponds to low-pass filtering of the spectrum of the heterodyned signal. Since the frequency ω_k is shifted to 0 before low-pass filtering, the filter effectively works as a band-pass filter centered at ω_k. The operations of Equations (18) and (19) are illustrated as a block diagram in Figure 10, which represents a single channel in the STFT filter-bank.

To obtain an analysis STFT which is equal to the one described by Equation (12), the sub-band signals must be downsampled. Figure 11 illustrates the downsampled STFT filter bank. Each sub-band signal is downsampled by the factor H_a, which corresponds to the hop size in the OLA interpretation. That is, each sub-band signal

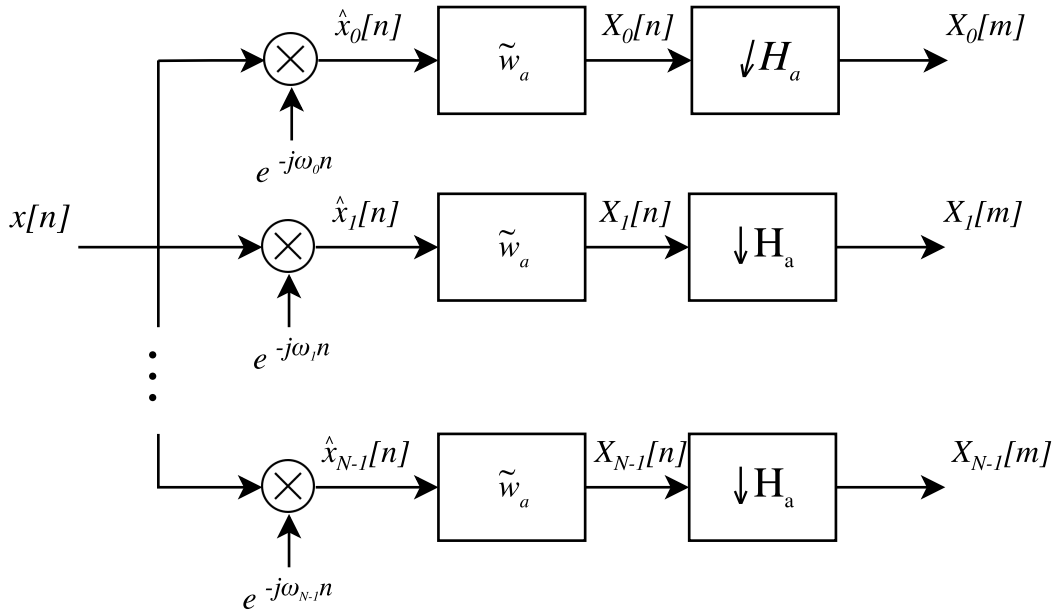


Figure 11: Downsampled STFT filter bank.

is evaluated at every H_ath sample. The downsampled STFT can be written as

$$\begin{aligned} X_k[m] &= \sum_{n=-\infty}^{\infty} \left(x[n]e^{-j\omega_k n}\right) w_a[n - mH_a] \\ &= \sum_{n=-\infty}^{\infty} \hat{x}_k[n] w_a[n - mH_a] \\ &= (\hat{x}_k * \tilde{w}_a)[m], \end{aligned} \tag{20}$$

where the mth sampling instant is at time $n = mH_a$.

3.1.2 Synthesis

First, let us consider the OLA interpretation of the STFT. In the synthesis, short time signals $y_m[n]$ are obtained by an inverse DFT on the synthesis frames $Y_m[k]$:

$$y_m[n] = \frac{1}{K} \sum_{k=0}^{K-1} Y_m[k] e^{j\omega_k n}. \tag{21}$$

The output signal can be reconstructed by summing all the short-time signals $y_m[n]$ [31]:

$$y[n] = \sum_{m \in \mathbb{Z}} y_m[n - mH_s] w_s[n - mH_a] \tag{22}$$

where $w_s[n]$ is the synthesis window. The synthesis frame rate is defined by the synthesis hop size H_s. That is, the mth time shifted short-time signal $y_m[n - mH_s]$ is centered around time $n = mH_s$. If no modifications are done before re-synthesis, that is, $Y[m, k] = X[m, k]$ and $H_s = H_a$, Equation (22) yields an output signal that

is equal to the input signal, given that the product of the analysis and synthesis windows satisfies the constant overlap-add (COLA) condition [34]:

$$\sum_{m\in\mathbb{Z}} w_a[n - mH_a]w_s[n - mH_s] = 1, \forall n \in \mathbb{Z} \tag{23}$$

A typical choice is to use the Hann window as both the analysis and synthesis windows. Then, the combined effect of the analysis and synthesis windows is the squared Hann window, which satisfies the COLA condition in Equation (23), when the synthesis hop size is set to any $H_s = M/(3+v)$, where v is a positive real number, given that the value of the fraction is a whole number [34].

It should be noted that a modified STFT $Y[m, k]$ may not correspond to the STFT of any signal $y[n]$. In this case, the reconstruction with Equation (22) yields an output signal $y[n]$ whose STFT is an approximation of $Y[m, k]$. In [25], an alternative reconstruction is proposed. It is based on setting the synthesis window equal with the analysis window, and scaling the short time signals $y_m[n]$ by the total power of the window functions:

$$y[n] = \frac{\sum_{m\in\mathbb{Z}} y_m[n - mH_s]w_s[n - mH_s]}{\sum_{m\in\mathbb{Z}} w_s^2[n - mH_s]}. \tag{24}$$

Using Equation (24) for reconstruction minimizes the mean squared error between the modified STFT $Y[m, k]$ and the STFT computed from the reconstructed signal $y[n]$.

Now, let us consider the FBS interpretation of the STFT synthesis stage. The synthesis filter bank is illustrated in Figure 12. Let us consider a synthesis STFT $Y[m, k]$, which is viewed as a frequency-ordered collection of narrow-band time-domain signals $Y_k[m]$. First, the narrow-band signals are upsampled according to the synthesis hop-size H_s. Next, the signals are convolved with the time-reversed synthesis window function $\tilde{w}_s$:

$$\hat{y}_k[n] = (Y_k * \tilde{w}_s)[n], \tag{25}$$

where $Y_k[n]$ are the upsampled sub-band signals. The output signal is the sum of the sub-band signals frequency shifted by ω_k:

$$y[n] = \sum_{k=0}^{K-1} \hat{y}_k[n]e^{j\omega_k n}. \tag{26}$$

Again, if no modifications are done before re-synthesis, that is, $Y[m, k] = X[m, k]$ and $H_s = H_a$, the output signal is equal to the input signal if the window functions satisfy the COLA condition in Equation (23).

3.2 Time-Scale Modification with the Phase Vocoder

TSM with the phase vocoder is based on two modifications to the STFT. First, a different hop size is used during analysis and synthesis, that is $H_a \neq H_s$. This

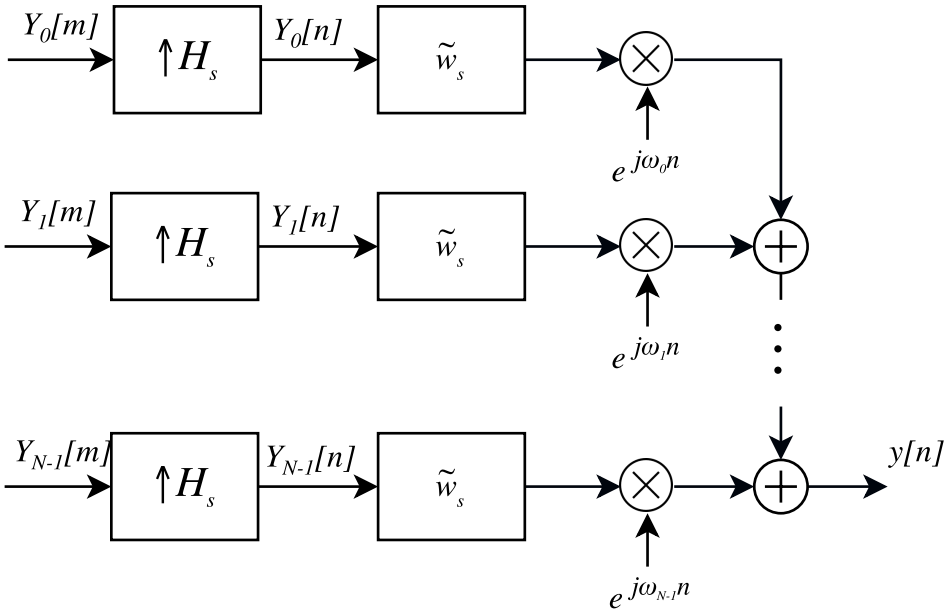


Figure 12: STFT synthesis filter bank.

means that the synthesis frames are relocated to a different time position than the corresponding analysis frames. Second, in order to preserve the periodicities in the input signal, the phases of the synthesis frames are modified. The phase modifications are done in a way that the periodic components in the time-shifted synthesis frames overlap coherently [15, 17, 36].

The modifications are based on a sinusoidal model of the input signal. The input signal is considered as a sum of $I[n]$ sinusoids, with time-varying amplitudes $A_i[n]$ and instantaneous frequencies ω_i [36, 17]:

$$x[n] = \sum_{i=1}^{I[n]} A_i[n] e^{j\phi_i[n]}, \text{ with} \tag{27}$$

$$\phi_i[n] = \phi_i[0] + \sum_{m=0}^{n} \omega_i[m], \tag{28}$$

where $\phi_i[n]$ is the instantaneous phase of the ith sinusoid, and $\omega_i[n]$ is the instantaneous frequency. Assuming a constant TSM-factor $\alpha = H_s/H_a$, the ideal synthesis phase of the ith sinusoid at time n would be:

$$\phi_s[n] = \phi_s[0] + \alpha \sum_{m=0}^{n} \omega_i[m], \tag{29}$$

where $\phi_s[0]$ is an arbitrary initial synthesis phase. Rearranging the terms in Equation (28), and inserting the solution to (29), the ideal synthesis phase can be written as:

$$\phi_s[n] = \phi_s[0] + \alpha\Big(\phi_i[n] - \phi_i[0]\Big). \tag{30}$$

The modifications done on the STFT representation in phase-vocoder TSM attempt to produce these ideal time-scaled sinusoids. The amplitudes of the time-scaled sinusoids are obtained by setting $|Y[m,k]| = |X[m,k]$. Since the hop size

during synthesis is different from the analysis hop size, the amplitudes of the sinusoids vary slower or faster, depending on the TSM factor, after re-synthesis of the output signal from the STFT representation.

To obtain the synthesis phases from the STFT representation, the instantaneous frequencies of the analysis STFT bins are estimated by:

$$\omega_{inst}[m,k] = \omega_k + \frac{1}{H_a}\kappa[m,k], \tag{31}$$

where $\kappa[m,k]$ is the estimated "heterodyned phase increment":

$$\kappa[m,k] = \Big[\angle X[m,k] - \angle X[m-1,k] - H_a\omega_k\Big]_{2\pi}. \tag{32}$$

Here, $\Big[\,\cdot\,\Big]_{2\pi}$ denotes the principal determination of the angle, i.e., the operator wraps the input angle to the range $[-\pi,\pi[$. An illustration of the estimation of the instantaneous frequency is shown in Figure 13, where the analysis phases are represented as unit vectors on the complex plane. The left panel shows the instantaneous phases for the bin k for two consecutive analysis frames. A rough estimate for the frequency of the sinusoid the bin is representing is given by the bin's center frequency. Based on the analysis phase of the previous frame $\angle X[m-1,k]$, the bin's center frequency ω_k, and the hop size H_a, a prediction is made that the instantaneous phase of the bin in the current frame is close to

$$\angle\tilde{X}[m,k] = \angle X[m-1,k] + H_a\omega_k. \tag{33}$$

Since the analysis phase of the current frame is known, the frequency estimate can be improved by taking into account the difference between the principal determination of the estimated instantaneous phase $\angle\tilde{X}[m,k]$, and the actual analysis phase $\angle X[m,k]$, which is the heterodyned phase increment described above.

The instantaneous frequency is an estimate of the rate of change of the analysis phase, i.e., the time derivative of the phase. Once the instantaneous frequencies have been estimated for the analysis bins, the synthesis phases can be computed by integrating the instantaneous frequencies over time, according to the synthesis hop size:

$$\angle Y[m,k] = \angle Y[m-1,k] + H_s\omega_{inst}[m,k], \tag{34}$$

where the discrete-time integration is done by summation.

3.2.1 Phase Coherence

In the phase vocoder, the input signal is modeled as a sum of sinusoids, such that each sinusoid corresponds to one frequency bin in the analysis STFT. The frequencies of the sinusoids are given by the estimated instantaneous frequencies. The phase propagation applied during TSM ensures that there are no phase discontinuities in these sinusoids between the time shifted synthesis frames. Thus, it is said that the phase vocoder ensures the "horizontal phase coherence" in the synthesis STFT. However, the phase relations across the bins in the frequency direction are lost in this process. This results in the loss of "vertical phase coherence".

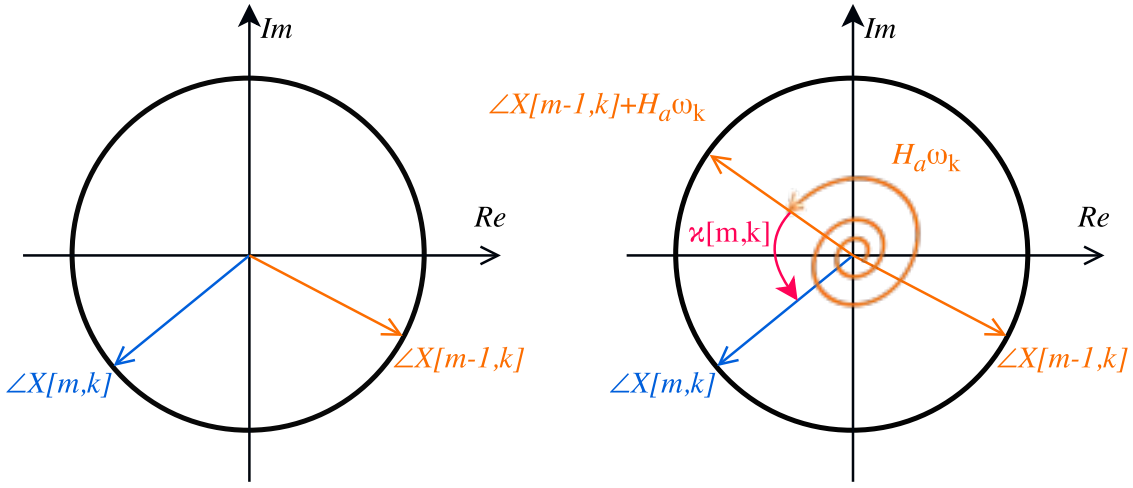
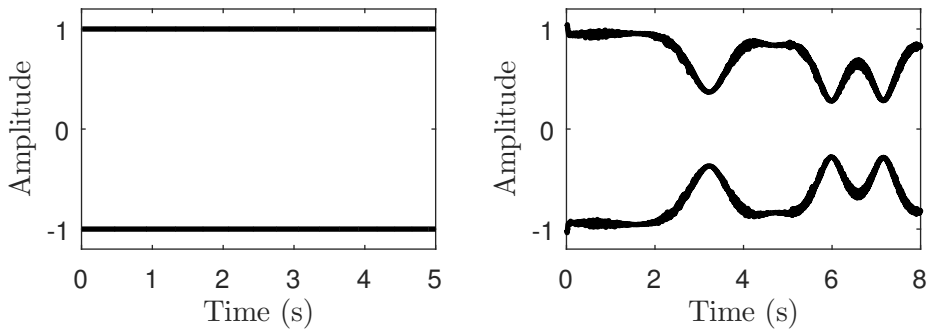


Figure 13: Estimation of the instantaneous frequency in the phase vocoder. The figure has been adapted from [3].

The loss of vertical phase coherence results in an artifact in the modified signal typically described as "phasiness", or as loss of clarity. For a constant-amplitude sinusoidal chirp signal, the loss of vertical phase coherence leads to fluctuations in the amplitude of the chirp. This is demonstrated in Figure 14, where the amplitude of a constant-amplitude sinusoid is shown after phase vocoder TSM with TSM-factor $\alpha = 1.6$. The amplitude of the sinusoid is dependent on the phase relationships of the bins representing the sinusoid in the synthesis STFT. As they are different in each synthesis frame, the amplitude of the chirp signal varies over time. The problem of phase coherence is discussed further in Section 4, where methods which attempt to preserve the vertical phase coherence are discussed.



(a) (b)

Figure 14: (**a**) The amplitude of a constant-amplitude chirp. (**b**) The amplitude of the constant-amplitude chirp after phase vocoder TSM with TSM factor $\alpha = 1.6$. The figure is adapted from [17].

3.2.2 Transient Smearing

Transients processed with the phase vocoder also suffer from the loss of vertical phase coherence. As the phase relationships across the frequency bins representing a transient are lost, the concentrated energy of the transient in the analysis frame gets smeared inside the synthesis frame after re-synthesis with the modified phase values. This is illustrated in Figure 15. An analysis frame $x_m[n]$, containing a transient is shown in Figure 15a. After applying phase propagation, the corresponding synthesis frame $y_m[n]$, which is shown in Figure 15b, is obtained. Due to loss of vertical phase coherence, the energy of the transient is distributed more evenly across the synthesis frame, and the impulsive nature of the transient is lost.

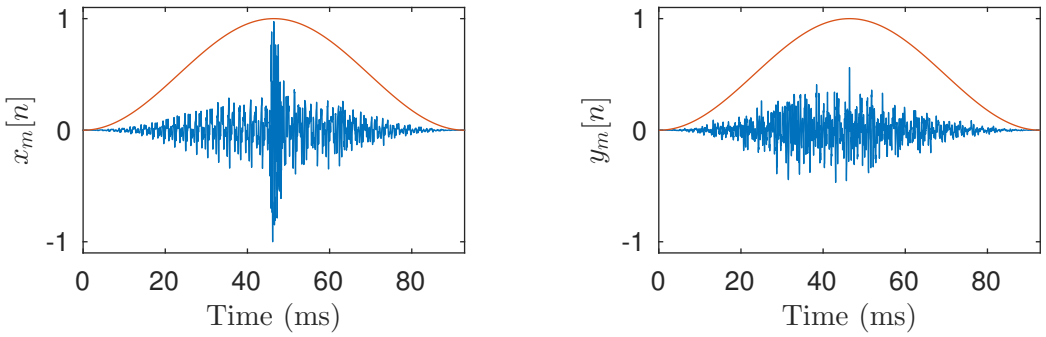


Figure 15: Transient smearing due to loss of vertical phase coherence in phase vocoder TSM. (**a**) An analysis frame containing a transient. (**b**) The corresponding synthesis frame which has the same magnitude spectrum as the analysis frame, but a different phase spectrum due to the phase modifications.

Furthermore, because a single transient is represented in several consecutive analysis frames, the transient energy also gets smeared across several synthesis frames. In the case of time expansion, in which the synthesis hop size is larger than the analysis hop size, this means that the transient's energy gets even further spread in time.

4 Phase Vocoder Extensions

In this section, some extensions to the standard phase vocoder based TSM are reviewed. Most of the methods aim in some way to improve the vertical phase coherence during modification. Other methods address the problems arising from the fixed time and frequency resolution of the STFT.

4.1 Intra-Sinusoidal Phase Locking

As explained in Section 3, the phase vocoder is based on a sinusoidal model of the input signal. Let us consider a constant-frequency sinusoid as the input signal:

$$x[n] = e^{j\omega n}. \tag{35}$$

Considering a time-limited analysis window of length N, its DFT is given by:

$$W_a[k] = \sum_{n=-N/2}^{N/2-1} w_a[n] e^{-j\omega_k n}. \tag{36}$$

Taking the DFT of the sinusoid in Equation (35), using the analysis window $w_a[n]$, yields:

$$X[k] = \sum_{n=-N/2}^{N/2-1} e^{j\omega n} w_a[n] e^{-j\omega_k n} = W_a[k - \frac{\omega}{2\pi} K], \tag{37}$$

where $K \geq N$ denotes the number of frequency bins in the DFT. The resulting spectrum is the DFT of the analysis window frequency shifted by the sinusoid's frequency ω. Thus, in the phase vocoder analysis, a constant frequency sinusoid excites not only one, but multiple frequency bins in each STFT frame. The number of excited bins depends on the bandwidth of the analysis window. If the analysis window is chosen as the Hann window, in the general case, a sinusoid excites four bins [16]. In the special case that the sinusoid's frequency is exactly tuned to a DFT bin's frequency, only three bins are excited.

Figure 16 shows a constant-frequency sinusoid windowed with three Hann windows, each with different phase characteristics. Figure 16a shows the regular Hann window shape, where the maximum is in the middle of the window, and the end points are at zero amplitude. Figure 16b shows the zero-phase Hann window, where the maxima of the window are at the edges, and there is a trough in the middle. Finally, Figure 16c shows a random-phase Hann window, which was obtained by taking the DFT of the regular Hann window, randomizing the phases in the frequency domain, and re-synthesizing the resulting spectrum back to time domain. The magnitude spectrum of these three signals are equal. They all have the magnitude spectrum of the Hann window frequency shifted by the sinusoid's frequency. The phase spectra are different however. In the case of Figure 16a, the phases of the excited bins are alternately π radians out of phase. In the case of Figure 16b, the excited bins are

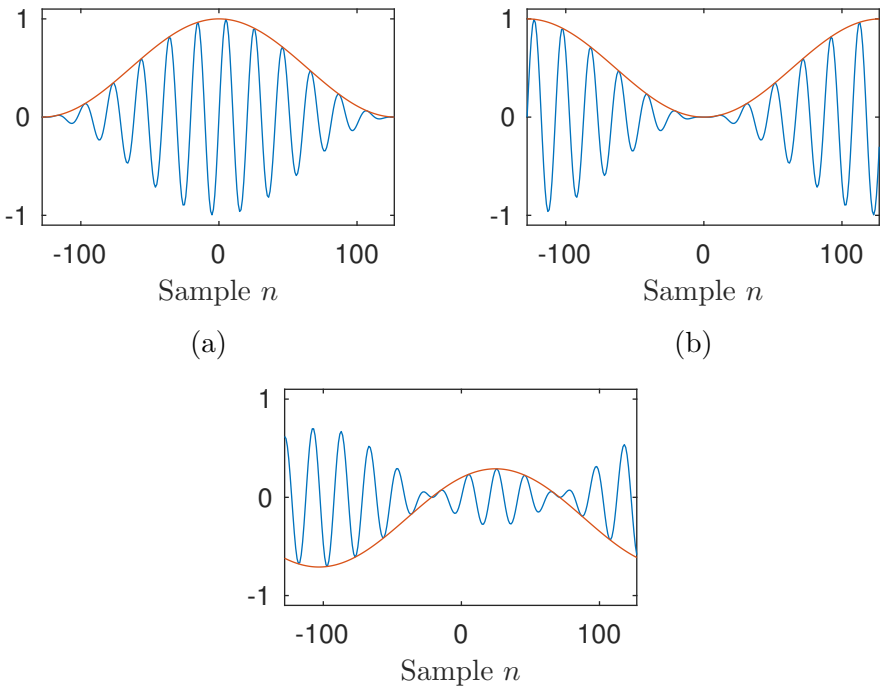


Figure 16: A constant-frequency sinusoid windowed with (**a**) the regular Hann window, (**b**) the zero-phase Hann window, and (**c**) a random-phase Hann window.

exactly in phase. Finally, as explained above, in Figure 16c, the phase relations between the bins are random.

To further illustrate the way sinusoids are represented in the STFT, Figure 17 shows the magnitude and phase spectrum of a zero-phase-windowed harmonic signal. Figure 17a shows the signal windowed by the linear-phase Hann window which reaches it maximum value at the center of the frame, and tapers to zero at the edges. To obtain a zero-phase spectral representation of the signal, samples inside the analysis frame are circularly shifted such that the samples in the latter half of the frame are located in the start of the frame and vice versa. This results in the signal shown in Figure 17b. The DFT is computed for this signal, which results in the magnitude and phase spectra shown in Figures 17c and 17d, respectively. It can be seen that the five sinusoids composing the harmonic signal each excite either three or four bins in the spectrum. As the signal is zero-phase windowed, the bins representing each sinusoid are exactly in phase. The excited bins are highlighted with orange lines.

Now consider the phase propagation applied in the standard phase-vocoder-based TSM, which was given in Equations (31–34). For each bin in the analysis STFT, the instantaneous frequencies are estimated, and the phases are modified accordingly to obtain the synthesis STFT. Therefore, the phases of the bins representing a single sinusoid are modified individually. If there are any errors in the estimation of the instantaneous frequencies, the phase relations of the bins in the synthesis STFT are

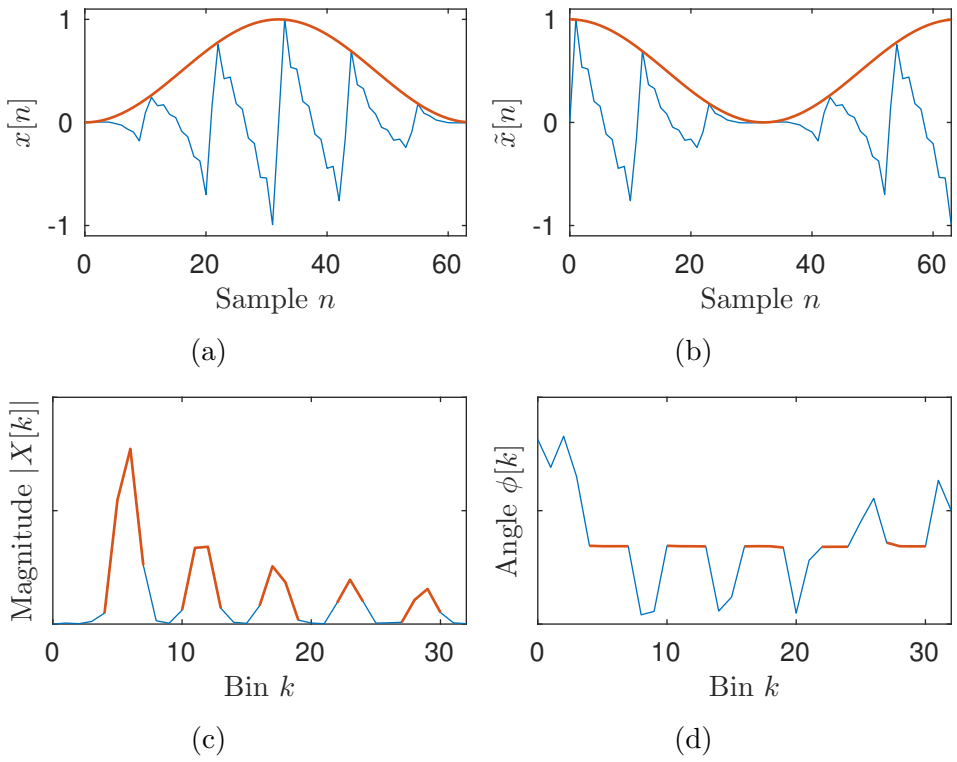


Figure 17: (**a**) A harmonic signal weighted by an analysis window. (**b**) A zero-phase version of the signal which was obtained by circularly shifting the samples such that the latter half of the signal moves to the start of the frame, and vice versa. (**c**) The magnitude spectrum of the signal. The bins representing the sinusoidal components of the signal are shown in orange. (**c**) The phase spectrum of the zero-phase signal. The phases of the sinusoidal bins are aligned, which appears as flat sections in the phase spectrum.

not equal to the ones in the analysis STFT. Also, in phase-vocoder-based TSM, the phase propagation is applied in an iterative fashion: the phases of the current synthesis frame depend not only on the current and previous analysis STFT frames, but also on the previous synthesis frame. Thus, even if the instantaneous frequencies of the bins are estimated correctly on the current frame, the phase relations of the previous synthesis frame have an effect on the resulting phases. If the original phase relations have been lost at any frame before the current one, the phases will not be aligned in any future frames either.

Thus, when the modified spectrum is synthesized back to time domain, the amplitude envelope of the synthesized signal does not necessarily have the shape of the original analysis window. For instance, it is possible that if the analyzed short-time signal corresponds to Figure 16a, the synthesized short-time signal resembles the one in Figure 16c, due to the applied phase modifications. Because the synthesized signal is no longer at zero amplitude in the edges, a synthesis window needs to be applied. The energy of the signal is no longer concentrated at the center and as such, applying the synthesis window results in an additional loss of energy. This leads to

fluctuations in the amplitude of the modified signal, since different short-time signals lose different amounts of energy when the synthesis window is applied, depending on how the energy is distributed in each short-time signal. This is the main cause of the phasiness artifact in the phase vocoder. To solve this problem, there needs to be a way to retain the phase relations of bins representing an individual sinusoid. This is referred to as preservation of the intra-sinusoidal phase coherence, or simply, phase locking.

4.1.1 Loose Phase Locking

The problem of intra-sinusoidal phase coherence was first addressed by Puckette [16]. The idea is that after the phases of the bins in the synthesis STFT have been computed using standard phase vocoder processing, a phase correction is then applied such that the phases of the neighboring bins affect each other. If zero-phase windowing is used in computing of the short-time spectra, the final synthesis phase for each bin is the phase of the complex number

$$Z[m,k] = Y[m,k] + Y[m,k-1] + Y[m,k+1], \tag{38}$$

where $Y[m,k]$ is the synthesis STFT computed with the standard phase vocoder. The phase of each bin is a weighted average of the phases of the bin itself and its two neighboring bins in the frequency direction. The weighting is determined by the magnitudes of the bins: when complex numbers are summed, the number with the largest magnitude affects the phase of the resulting number the most. Thus, considering a sinusoidal signal in a synthesis STFT frame which affects three adjacent bins, where the amplitude of the middle bin is the largest, the phases of the bins with the smaller magnitude approach the phase of the middle bin.

4.1.2 Rigid Phase Locking

Alternative phase locking schemes were proposed by Laroche and Dolson [17, 36]. As opposed to the method of Puckette [16], in which the same processing is applied to each bin, these methods are based on finding peaks in the spectrum, and processing the bins related to each peak in a joint fashion. Peaks in the spectrum are defined as bins whose amplitude is larger than its four closest bins in the frequency direction. For each peak bin, the surrounding bins are attached to the bin's "region of influence", and as such, the frequency axis in each short-time spectra is subdivided into these regions. The upper limit of the region of a peak is chosen as the bin whose frequency is closest to the average frequency of that peak bin and the next one. Alternatively, the bin with the lowest magnitude between two consecutive peaks can be chosen as the boundary between two bin regions.

The idea is then to apply standard phase vocoder processing only to the peak bins, and to somehow lock the phases of the remaining bins to the phase of their corresponding peak bin. Two phase locking schemes are introduced: 1) identity phase locking, and 2) scaled phase locking. In identity phase locking, the idea is that for each peak bin and the surrounding bins, the phase relations between the bins

in the synthesis STFT are equal to the ones in the analysis STFT. Given a peak bin $Y[m, k_p]$ in the synthesis STFT whose phase has been computed according to Equations (31–34), the phases of the bins in its region of influence are updated as:

$$\angle Y[m, k] = \angle X[m, k] + \Big[\angle Y[m, k_p] - \angle X[m, k_p]\Big]_{2\pi}, \tag{39}$$

where $X[m, k]$ is the analysis STFT. Thus, for each non-peak bin, the phase update consists only of adding a value to its analysis phase. The value added is the principal determination of the phase difference between the corresponding peak bin's synthesis and analysis phases.

Scaled phase locking is an extension of the identity phase locking scheme, in which peak trajectories are formed between peaks in consecutive frames. Consider a sinusoid whose frequency varies such that its peak bin moves from bin k_0 to bin k_1 between the previous and current frame. In this case the phase integration in Equation (34) should be modified such that

$$\angle Y[m, k_1] = \angle Y[m - 1, k_0] + H_s \omega_{inst}[m, k_1], \tag{40}$$

where the phase increment is added to the phase of the peak bin in the previous frame, rather than the phase of the current peak bin in the previous frame. For each peak bin in the current frame, the corresponding peak bin in the previous frame is determined by finding the region of influence the current peak bin is in the previous frame, and finding the peak bin of that region.

Given a peak bin in the current frame k_1 and a corresponding peak bin in the previous frame k_0, the heterodyned phase increment is then computed as:

$$\kappa[m, k_1] = \Big[\angle X[m, k_1] - \angle X[m - 1, k_0] - \frac{H_a}{2}(\omega_{k_1} + \omega_{k_0})\Big]_{2\pi}, \tag{41}$$

where the average of the bin frequencies of the two peak bins are used. The instantaneous frequency is given by:

$$\omega_{inst}[m, k_1] = \frac{\omega_{k_1} + \omega_{k_0}}{2} + \frac{1}{H_a}\kappa[m, k_1], \tag{42}$$

and the synthesis phase of the peak bin is then computed according to Equation (40). The phases of the non-peak bins are locked to the peak bin phase by

$$\angle Y[m, k] = \angle Y[m, k_1] + \beta\Big[\angle X[m, k] - \angle X[m, k_1]\Big]_{2\pi}, \tag{43}$$

where β is a phase scaling factor. Setting $\beta = 1$ corresponds to identity phase locking. Alternative choices of β are also discussed, but no definite "correct" choice is given. However, it is stated that the value should be between 1 and the TSM factor α.

4.2 Transient Preservation

The assumption in the way the phase modifications are done in the phase vocoder is that the input sound can be represented as a sum of slowly varying sinusoids. In the

case that there is an abrupt increase in energy, that is, a transient in the input signal, this assumption is not valid. Therefore, when standard phase vocoder processing is applied to transients, the results are often of poor quality. The loss of quality can be attributed to the loss of vertical phase coherence between frequency bins which represent the transient. Because the phase relations between the bins are modified, the synthesized transient does not resemble the original analyzed transient.

To overcome this limitation, several solutions which attempt in some way to preserve the vertical phase coherence of bins related to transients in the input signal have been proposed. This means that a way to detect transients is needed. A natural solution for preserving the vertical phase coherence is to apply a phase reset at detected transient locations of the input signal. That is, when a transient is detected, the phases of the bins in the analysis STFT are directly used in the synthesis STFT. This approach was originally introduced by Quatieri et al. [37]. In this method, whenever a local maximum (in the time direction) is detected at a STFT bin, a phase reset is applied to that bin. The method relies on the assumption that a transient appears as local maxima at several bins simultaneously to preserve the vertical phase coherence.

4.2.1 Vertical Phase Coherence at Transients

The concept of applying a phase reset at transients was further explored by Bonada [38]. First, it must be noted that the method differs from the standard phase vocoder based TSM in that the same hop size is used during analysis and synthesis. Figure 18 illustrates the approach used. When a TSM factor $\alpha > 1$ is used, that is, the time scale of the signal is expanded, some analysis frames are used twice during the synthesis. In the case of time contraction $\alpha < 1$, some analysis frames are omitted from the synthesis. This approach can be considered as a time-varying TSM factor, which changes locally whenever analysis frames are either reused or omitted.

Because the TSM factor is varied locally, it is possible to set the TSM factor to one during transients, similar to the approach in the SOLA extensions [21, 22] presented in Section 2. Whenever a transient is detected, all the analysis frames are used exactly once in the synthesis. The local change in the TSM factor needs to be compensated in the steady state regions of the sound, such that the average TSM factor corresponds to the desired amount of TSM. Transients are detected by considering the changes in the STFT bin energies, and the changes in the computed Mel cepstrum coefficients. The transient detection stage can distinguish between transients which only affect the energy in the high frequencies, low frequencies or in the whole band. According to the type of the transient, a phase reset is applied to bins in the region of influence of the transient, while standard phase modifications are applied to the remaining bins.

A similar approach for transient preservation was proposed by Duxbury et al. [39]. It is also based on locally changing the TSM factor to one during detected transients, and compensating for it in the steady state regions. The transients are detected from changes in the estimated transient energy. The transient energy is estimated by first splitting the input signal to a steady state component, and a residual component,

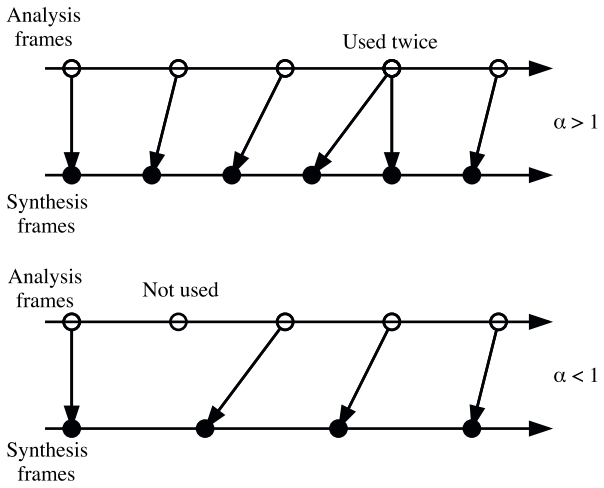


Figure 18: Analysis and synthesis frames used in the method of Bonada [38], where the analysis frames are either used more than once, or omitted completely during synthesis, depending on whether time expansion or time contraction is desired, respectively. The figure has been adapted from [38].

which contains the transients and noise, using a multi-resolution technique presented in [40]. The transients are detected as local maxima in the time derivative of the low-pass-filtered transient energy function. For the first frame in a set of subsequent transient frames, the heterodyned phase increment is computed as

$$\kappa[m,k] = \Big[\angle X[m,k] - \angle Y[m-1,k] - H_s\omega_k\Big]_{2\pi}, \tag{44}$$

where the difference between the current analysis phase $\angle X[m,k]$ and the predicted synthesis phase $\angle Y[m-1,k] + H_s\omega_k$ is used instead of the standard way, where the predicted analysis phase $\angle X[m-1,k] + H_a\omega_k$ is compared to the actual analysis phase. Apart from the estimation of the heterodyned phase increment, standard phase propagation is applied as in Equations (31, 34). This preserves the vertical coherence of the synthesis STFT bins such that the phase relations between the synthesis frame bins are equal to the ones in the analysis STFT, though the exact phase values are not.

The above methods suffer from the need to set the TSM factor to one during transients. Furthermore, these methods preserve the vertical phase coherence in relatively wide frequency bands [38] or for the whole audio band [39]. A method which addresses these limitations was proposed by Röbel [19]. In this method, the transient detection and preservation are done on the level of spectral peaks. The transients are detected as soon as the analysis window slides over them. A transient which is located on the edge of the window can be detected by the center of gravity (COG) of the analysis frame:

$$n_{cog}[m] = \frac{\sum_{n=-N/2}^{N/2-1} n x_m^2[n]}{\sum_{n=-N/2}^{N/2-1} x_m^2[n]}, \tag{45}$$

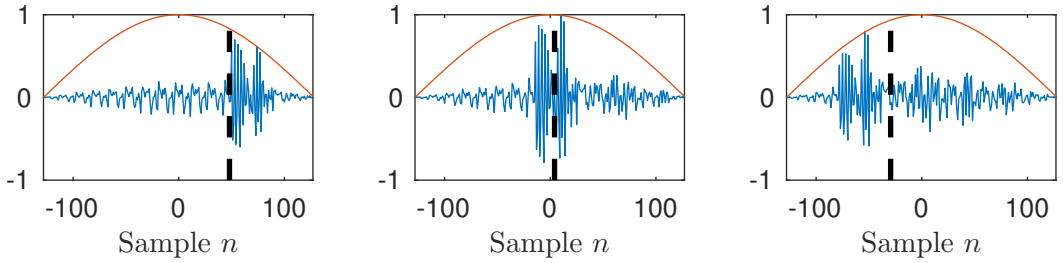


Figure 19: The COG of the analysis frame (marked with the black dashed line) when the transient is at (**a**) the right edge, (**b**) the center, and (**c**) the left edge of the analysis window.

where $x_m[n]$ is the mth analysis frame of length N. Figure 19 shows the COG of the short-time signal as the analysis window slides over a transient. A large positive value of n_{cog} indicates that the analysis window has slid over a transient (Figure 19a). A frame COG value close to zero indicates that the transient is close to the center of the analysis frame (Figure 19b), whereas a negative n_{cog} value indicates that the analysis window has slid over the transient (Figure 19c). Thus, COG values above a certain threshold can be used for transient detection.

In order to detect transient on a level of spectral peaks, the computation of the COG can be performed using a subset of the STFT frequency bins:

$$n_{cog} = \frac{\sum_{k=k_l}^{k_h} -\frac{\partial \angle |X[m,k]|}{\partial k} |X[m,k]|^2}{\sum_{k=k_l}^{k_h} |X[m,k]|^2}, \tag{46}$$

where k_l and k_h are the lowest and highest frequency bins in the region of influence of the spectral peak, respectively. The robustness of the transient detection is improved by a statistical model, which attempts to distinguish between positive COG values introduced by transients, from ones introduced by amplitude modulated sinusoids or noise.

Transients detected at the level of spectral peaks are further grouped on a sub-band level. When a transient onset is detected on a sub-band, the spectral peaks whose COG values are above a threshold are collected into a non-contracting set of transient bins K_t, until the end of the transient event is detected. In order to prevent transient smearing, during synthesis, the magnitude and instantaneous frequencies of the previous frame are used for the transient bins. When the end of the transient event is detected, that is, when the analysis window is centered on the transient, the phases of the transient bins are reset. Since the frames preceding the transient frame do not contribute to the transient‘s energy, the magnitude of the transient bins is compensated by multiplying them with 1.5.

4.2.2 Harmonic and Percussive Separation

In [41], a method for transient preservation which is not based on phase resets or phase locking was proposed. Figure 20 illustrates the method. The input signal is first

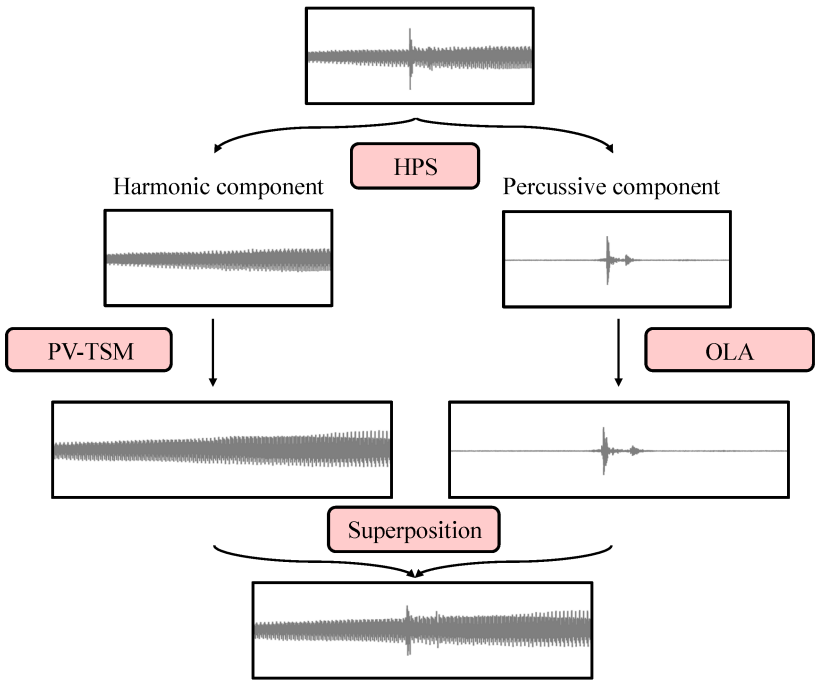
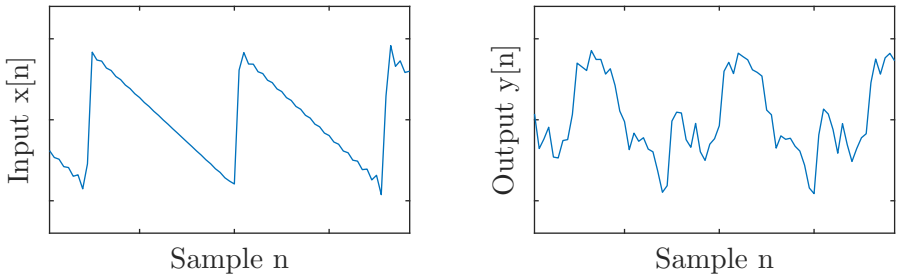


Figure 20: Time-scale modification (TSM) based on harmonic and percussive separation. The figure has been adopted from [3].

separated into harmonic and percussive components. The harmonic and percussive separation is based on the median filtering technique proposed by Fitzgerald [42]. Due to the distinct nature of the separated signal components, the harmonic and percussive signals are processed using different techniques. Standard phase vocoder processing, using a relatively long analysis window is used for the harmonic signal, whereas standard OLA processing is used for the percussive signal. The output is the superposition of the two separately processed signals. It was shown that standard OLA processing with a short enough window provided surprisingly good TSM quality when applied to transient signals. The method is appealing in that it can preserve transients during TSM without the need to detect individual transient locations, which is considered a difficult and error-prone task. Furthermore, it allows simultaneously using an analysis window with a good frequency resolution for the harmonic signal, and an analysis window with a good time resolution for the percussive signal.

4.3 Shape-Invariance

The various phase-locking schemes reviewed in Section 4.1 considered the preservation of the vertical phase coherence for individual sinusoids in the input signal. The methods for transient processing reviewed in Section 4.2 considered the preservation of the vertical phase coherence for bins representing a transient in the input signal. What these methods do not address, however, is the preservation of the phase relations between the partials of a harmonic signal. The loss of inter-sinusoidal phase



(a) (b)

Figure 21: (**a**) The input sawtooth waveform and (**b**) the sawtooth modified with the phase-locked vocoder.

coherence changes the waveform of the processed harmonic signal. Figure 21 shows a sawtooth waveform which has been processed with the phase-locked vocoder using identity phase locking [17], such that the intra-sinusoidal phase coherence has been preserved but the inter-sinusoidal phase coherence has not been addressed. It can be seen that the original shape of the sawtooth is not preserved during TSM. While the loss of the shape of the time-domain waveform is uncritical for most signals, it is clearly audible for speech, because the perception of the underlying excitation pulses is affected.

This problem has been addressed by introducing "shape-invariant processing" for the phase vocoder. A solution was first proposed by Quatieri and McAulay in the context of sinusoidal modeling of speech signals [43]. Röbel applied the shape-invariant property for phase vocoder based speech transformation [44]. The phase propagation in the shape-invariant phase vocoder is given by:

$$\angle Y[m,k] = \angle X[m,k] + (\kappa[m,k] + \frac{2\pi k}{K})\Delta_m, \tag{47}$$

where Δ_m is the time shift which has to be applied to the synthesis frame in order to maximize alignment with the preceding synthesis frame. Similarly to the time-domain techniques SOLA and WSOLA, the optimal time shift is determined by cross-correlation. In the shape-invariant phase vocoder, the cross-correlation is computed for two consecutive synthesis frames in the frequency domain by:

$$r_m[\Delta] = \sum_{k=0}^{K} Y_m^*[k] Y_{m-1}[k] e^{j\omega_k n}, \tag{48}$$

where K is the number of bins in each short-time spectra, $Y_m^*[k]$ is the complex conjugate of the current synthesis frame, and Y_{m-1} is the previous synthesis frame. To ensure that the estimated optimal time shift is only affected by the tonal components of the input signal, a sinusoidal mask can be applied to the synthesis frames prior to the computation of the cross correlation sequence. The optimal time shift Δ_m is searched from the cross-correlation sequence $r_m[\Delta]$ by finding the time shift which maximizes its value after taking into account the effect of the analysis windows in the synthesis frames.

4.4 Other Extensions

4.4.1 Sinusoidal Modeling

An important technique related to the phase vocoder is sinusoidal modeling, which was originally introduced by Quatieri and McAulay [45, 46]. In sinusoidal modeling, the time-frequency representation of the input signal which is obtained by phase vocoder analysis, is explicitly modeled as a sum of time-varying sinusoids. For each short-time spectra, the most dominant peaks are estimated. Then, for each peak, the amplitudes, frequencies and phases of sinusoids which correspond to these peaks are estimated. Additionally, sinusoidal peak continuation is enforced by forming trajectories between sinusoidal peaks in consecutive frames. After estimating the sinusoidal parameters and forming the trajectories, the modeled signal is obtained by additive synthesis of the estimated sinusoids.

The sinusoidal model was later extended with the sinusoids and noise model by Serra [47]. In the sinusoids and noise model, the part of the input signal which is not modeled by the time-varying sinusoids, known as the residual, is modeled as noise. The residual can be computed by subtracting the signal which was given by the sinusoidal model from the original signal. This can be done on a frame-by-frame basis, which results in a series of short-time spectra which represent the noise component. The short-time noise spectra are modeled by estimating their spectral envelopes. The modeled residual can be re-synthesized to the time domain from the short-time spectra by applying random phases to all the spectral bins and using standard phase vocoder synthesis on the resulting representation.

Because the sinusoids and noise model represents the input signal as a sum of slowly-varying sinusoids and noise, it is unable to preserve the quality of transients in the input signal. Thus, the model was extended by separate transient modeling, which resulted in the sinusoids, transients and noise model [48, 49]. In the method, the transients are first extracted from the input signal, and the remainder is modeled by sinusoids and noise. During TSM, the individual transients remain unmodified, and are only shifted in time according to the desired modification factor. The TSM for the sinusoids and noise is achieved by modifying the rate at which the sinusoidal and noise parameters evolve over time [50].

4.4.2 Multiresolution Techniques

Several techniques have been proposed which aim to alleviate the problems arising from the fixed time and frequency resolution of standard phase vocoder analysis. In [51], a technique based on the theory of nonstationary Gabor frames [52, 53] is proposed. In this technique, short-time spectra at each analysis time-instant are computed using multiple analysis window sizes, resulting in analyses with different time and frequency resolutions. Then, for each analysis time instant, the short-time spectra which gives the optimal representation of the underlying short-time signal is selected for the final time-frequency representation. The resolution is automatically adapted using an entropy-based sparsity criterium, which is based on the Rényi entropies [54]. A re-synthesis procedure from the adapted analysis coefficients is

also proposed. It is shown that for the problem of TSM, the adaptive method provides higher quality transformation when compared to results obtained with a fixed resolution, even when the fixed resolution is chosen optimally for each test signal. A different technique based on the theory of nonstationary Gabor frames is proposed in [55].

In [56], a multiresolution TSM technique is proposed, which is based on the multi-scale STFT [57]. In the multi-scale STFT, the input signal is first decomposed into multiple layers of various levels of "transientness". Then, the STFT is computed for each layer separately, such that the length of the analysis window is adjusted according to the transientness of the layer. That is, the analysis window is shorter for the more transient layers. In the TSM technique, standard phase propagation is first applied on the analysis STFT of the whole input signal. Next, the obtained synthesis phase values are applied to the STFTs computed for the different layers. Finally, the output signal is re-synthesized by applying standard phase vocoder synthesis for each layer separately and summing the signals together.

4.4.3 Partial Phase Derivatives

In the standard phase propagation procedure of the phase vocoder, as shown in Equations (31, 32, 34), the phase is first differentiated in the time direction to find the instantaneous frequencies, and then integrated according to the TSM-factor to compute the synthesis phases. Thus, the standard phase propagation does not take into account the partial phase derivative in the frequency direction in any way. In [58], a technique is developed in which the phase propagation is based on the full phase gradient estimated during the analysis stage. During the synthesis, the phase is propagated using the real-time phase gradient heap integration algorithm [59] which is able to take into account the partial phase derivative in the frequency direction. It was shown that the technique is able to preserve the quality of the transients during modification, even though no special handling of transients is done.

5 A Novel Time-Scale Modification Technique

In this section, a novel phase vocoder based TSM technique is proposed in which the applied phase propagation is based on the characteristics of the input audio. The input audio characteristics are quantified by means of fuzzy classification of spectral bins into sinusoids, noise, and transients. The information about the nature of the spectral bins is used for preserving the intra-sinusoidal phase coherence of the tonal components, while simultaneously preserving the noise characteristics of the input audio. Furthermore, a novel method for transient detection and preservation based on the classified bins is proposed. Most of the material in this section and in Section 6 has recently been published in [60].

5.1 Fuzzy Classification of Bins in the Spectrogram

The proposed method for the classification of spectral bins is based on the observation that, in a time-frequency representation of a signal, stationary tonal components appear as ridges in the time direction, whereas transient components appear as ridges in the frequency direction [42, 61]. Thus, if a spectral bin contributes to the forming of a time-direction ridge, most of its energy is likely to have originated from a tonal component in the input signal. Similarly, if a spectral bin contributes to the forming of a frequency-direction ridge, most of its energy is probably from a transient component. As a time-frequency representation, the STFT is used:

$$X[m,k] = \sum_{n=-N/2}^{N/2-1} x[n+mH_a]w[n]e^{-j\omega_k n}, \tag{49}$$

where $x[n]$ is the input signal, $H_a[a]$ is the analysis hop size, $w[n]$ is the analysis window, N is the analysis frame length and the number of frequency bins in each frame, and $\omega_k = \frac{2\pi k}{N}$ is the normalized center frequency of the kth STFT bin. Figure 22 shows the STFT magnitude of a signal consisting of a melody played on the piano, accompanied by soft percussion and a double bass. The time-direction ridges introduced by the harmonic instruments, and the frequency-direction ridges introduced by the percussion are apparent on the spectrogram.

The median filtering technique proposed by Fitzgerald [42] is used to compute the tonal and transient STFTs $X_s[m,k]$ and $X_t[m,k]$, respectively:

$$X_s[m,k] = \text{median}(|X[m-\frac{L_t}{2}+1,k]|,...,|X[m+\frac{L_t}{2},k]|) \tag{50}$$

and

$$X_t[m,k] = \text{median}(|X[m,k-\frac{L_f}{2}+1]|,...,|X[m,k+\frac{L_f}{2}]|), \tag{51}$$

where L_t and L_f are the lengths of the median filters in time and frequency directions, respectively. For the tonal STFT, the subscript s (denoting sinusoidal) is used and for the transient STFT the subscript t. Median filtering in the time-direction suppresses the effect of transients in the STFT magnitude, while preserving most of the energy

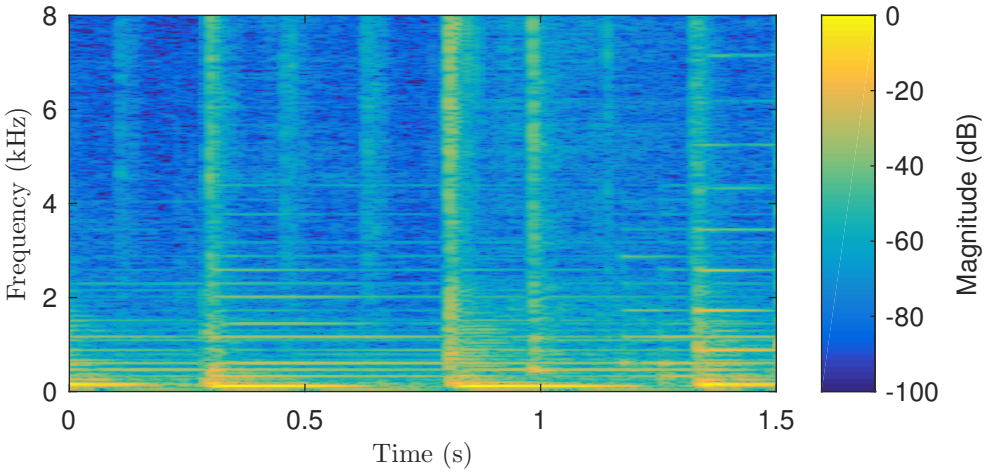


Figure 22: Spectrogram of a signal consisting of piano, percussion and double bass.

of the tonal components. Conversely, median filtering in the frequency-direction suppresses the effect of tonal components, while preserving most of the transient energy [42].

The two computed STFTs are used to estimate the tonalness, noisiness, and transientness of each analysis STFT bin. We estimate tonalness by the ratio

$$R_s[m,k] = \frac{X_s[m,k]}{X_s[m,k] + X_t[m,k]}. \tag{52}$$

We define transientness as the complement of tonalness:

$$R_t[m,k] = 1 - R_s[m,k] = \frac{X_t[m,k]}{X_s[m,k] + X_t[m,k]}. \tag{53}$$

Signal components which are neither tonal nor transient, can be assumed to be noiselike. Experiments on noise signal analysis using the above median filtering method show that the tonalness value is often approximately $R_s = 0.5$. This is demonstrated in Figure 23, where a histograms of the tonalness values of STFT bins of pink noise and white noise signals signal are shown. It can be seen, that the tonalness values are approximately normally distributed around the value 0.5. Thus, we estimate noisiness by

$$R_n[m,k] = 1 - |R_s[m,k] - R_t[m,k]| = \begin{cases} 2R_s[m,k], & \text{if } R_s[m,k] \leq 0.5 \\ 2(1 - R_s[m,k]), & \text{otherwise.} \end{cases} \tag{54}$$

The tonalness, noisiness, and transientness can be used to denote the degree of membership of each STFT bin to the corresponding class in a fuzzy manner. The relations between the classes are visualized in Figure 24.

Figure 25 shows the computed tonalness, noisiness, and transientness values for the STFT bins of the example audio used above. The tonalness values are close to 1

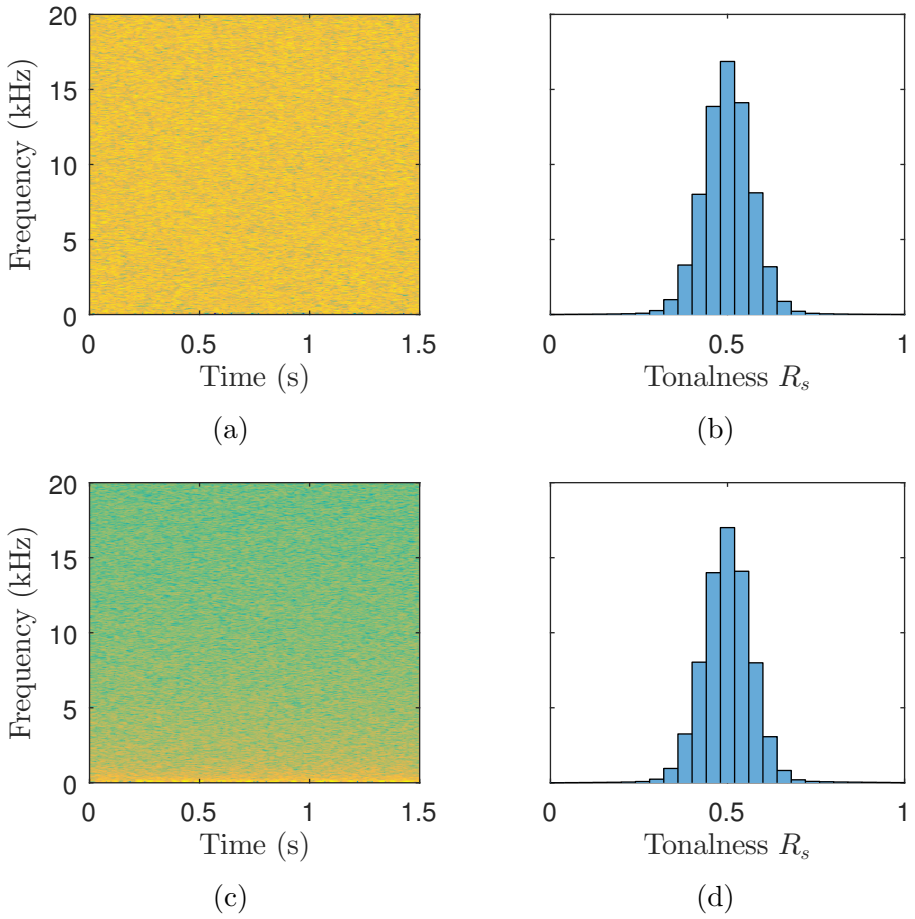


Figure 23: Spectrograms of (**a**) white noise and (**c**) pink noise. (**b**, **d**) show the histograms of their tonalness values.

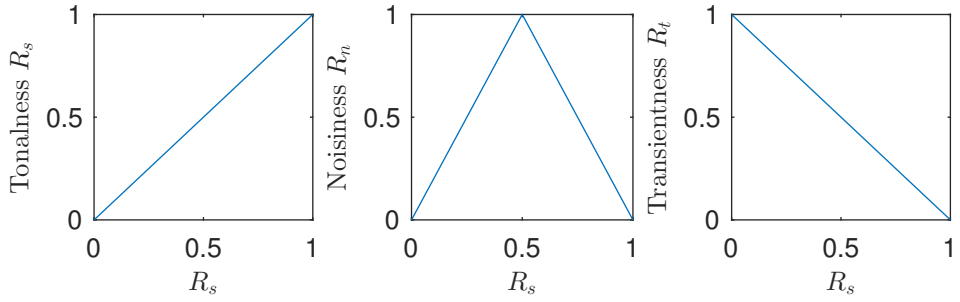
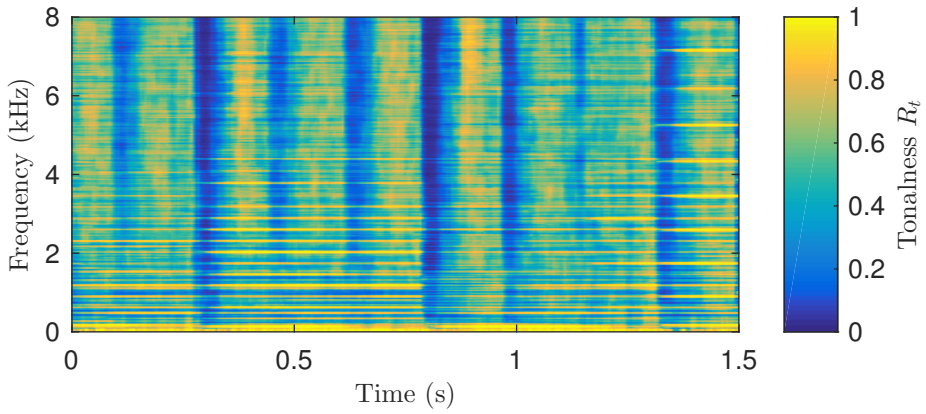
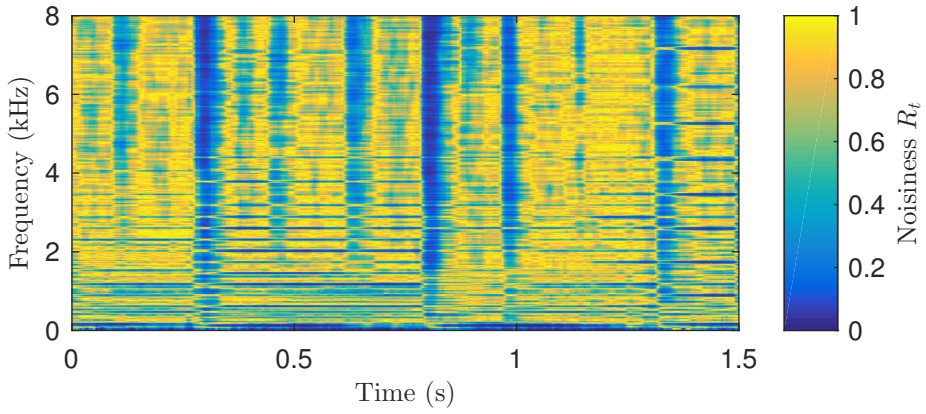


Figure 24: The relations between the fuzzy classes.

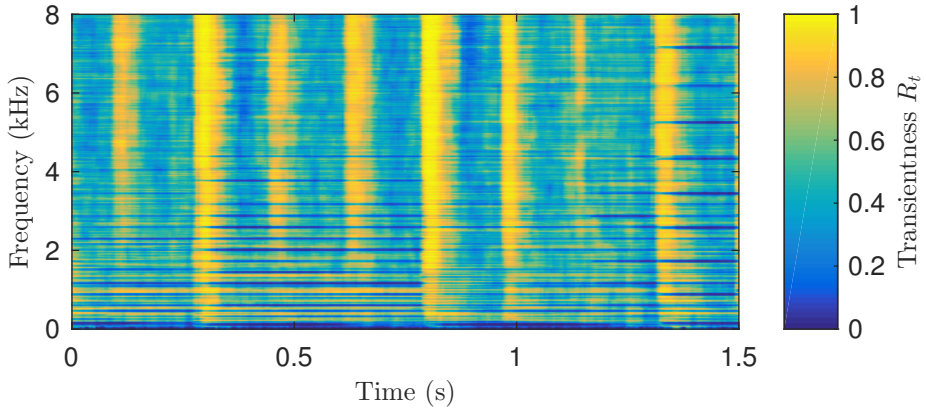
for bins which represent the harmonics of the piano and double bass tones, whereas the tonalness values are close to 0 for bins which represent the percussion hits. The noisiness values are close to 1 for bins which do not significantly contribute to the representation of either tonal or transient components in the input audio. Finally, it can be seen that the transientness values are complementary to the tonalness values.



(a)



(b)



(c)

Figure 25: (**a**) Tonalness, (**b**) noisiness, and (**c**) transientness values for the STFT bins of the example audio signal. Cf. Figure 22.

5.2 Time-Scale Modification Technique

This section introduces the new TSM technique that is based on the fuzzy classification of spectral bins defined above.

5.2.1 Proposed Phase Propagation

As discussed in Section 3, phase vocoder TSM is based on the differentiation and subsequent integration of the analysis STFT phases in time. This process is known as phase propagation. The phase propagation in the new TSM method is based on a modification to the phase-locked vocoder by Laroche and Dolson [17], which was reviewed in Section 4.1.2. For clarity, a brief description of the phase-locked vocoder is included here also. The phase propagation in the phase-locked vocoder can be described as follows. For each frame in the analysis STFT (49), peaks are identified. Peaks are defined as spectral bins, whose magnitude is greater than the magnitude of its four closest neighboring bins in the frequency direction.

The phases of the peak bins are differentiated to obtain the instantaneous frequency for each peak bin:

$$\omega_{inst}[m,k] = \omega_k + \frac{1}{H_a}\kappa[m,k], \tag{55}$$

where $\kappa[m,k]$ is the estimated "heterodyned phase increment":

$$\kappa[m,k] = \Big[\angle X[m,k] - \angle X[m-1,k] - H_a\omega_k\Big]_{2\pi}. \tag{56}$$

Here, $\Big[\,\cdot\,\Big]_{2\pi}$ denotes the principal determination of the angle, i.e., the operator wraps the input angle to the range $[-\pi,\pi[$. The phases of the peak bins in the synthesis STFT $Y[m,k]$ can be computed by integrating the estimated instantaneous frequencies according to the synthesis hop size:

$$\angle Y[m,k] = \angle Y[m-1,k] + H_s\omega_{inst}[m,k], \tag{57}$$

where H_s is the synthesis hop size. The ratio between the analysis and synthesis hop sizes determines the TSM factor $\alpha = H_s/H_a$. In the standard phase vocoder TSM [15], this kind of phase propagation is applied to all bins, not only peak bins. In the phase-locked vocoder [17], the way the phases of non-peak bins are modified is known as phase locking. It is based on the idea that the phase relations between all spectral bins, which contribute to the representation of a single sinusoid, should be preserved when the phases are modified. This is achieved by modifying the phases of the STFT bins surrounding each peak such that the phase relations between the peak and the surrounding bins are preserved from the analysis STFT. Given a peak bin k_p, the phases of the bins surrounding the peak are modified by:

$$\angle Y[m,k] = \angle X[m,k] + \Big[\angle Y[m,k_p] - \angle X[m,k_p]\Big]_{2\pi}, \tag{58}$$

where $\angle Y[m,k_p]$ is computed according to (55–57). This approach is known as identity phase locking.

From the motivation behind phase locking, it seems that it should only be applied to bins that are considered to represent a sinusoidal component in the input signal. When applied to non-sinusoidal bins, phase locking introduces a metallic sounding artifact to the processed sound. Since the tonalness, noisiness, and transientness

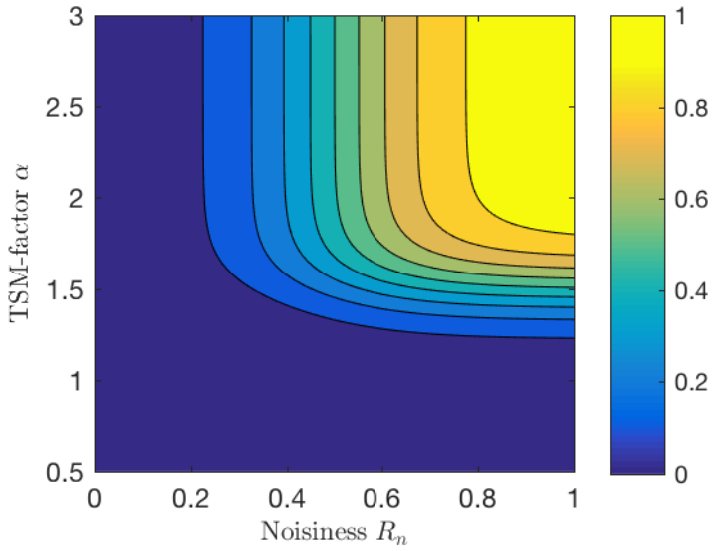


Figure 26: A contour plot of the phase randomization factor A_n, with $b_n = b_\alpha = 4$.

of each bin has been determined, this information can be used when phase locking is applied. We want to be able to apply phase locking to bins which represent a tonal component, while preserving the randomized phase relationships of bins representing noise. Thus, phase locking is first applied to all the bins. Secondly, phase randomization is applied to the bins according to the estimated noisiness values. The final synthesis phases are obtained by adding uniformly distributed noise to the synthesis phases computed with the phase-locked vocoder:

$$\angle Y'[m,k] = \angle Y[m,k] + \pi A_n[m,k](u[m,k] - \frac{1}{2}), \tag{59}$$

where $u[m,k]$ are the added noise values and $\angle Y[m,k]$ are the synthesis phases computed with the phase-locked vocoder. The pseudo-random numbers $u[m,k]$ are drawn from the uniform distribution $\mathcal{U}(0,1)$. $A_n[m,k]$ is the phase randomization factor, which is based on the estimated noisiness of the bin $R_n[m,k]$ and the TSM factor α:

$$A_n[m,k] = \frac{1}{4}\Big[\tanh(b_n(R_n[m,k]-1))+1\Big]\Big[\tanh(b_\alpha(\alpha - \frac{3}{2}))+1\Big], \tag{60}$$

where constants b_n and b_α control the shape of non-linear mappings of the hyperbolic tangents. The values $b_n = b_\alpha = 4$ were used in this implementation. The phase randomization factor A_n, as a function of the estimated noisiness R_n and the TSM factor α, is shown in Figure 26. The phase randomization factor increases with increasing TSM factor and noisiness. The phase randomization factor saturates as the values increase, such that at most, the uniform noise added to the phases gets values in the range $[-0.5\pi, 0.5\pi]$.

5.2.2 Transient Detection and Preservation

For transient detection and preservation, a similar strategy is adopted as in [19]. However, the proposed method is based on the estimated transientness of the STFT bins. Using the measure for transientness, the smearing of both the transient onsets and offsets is prevented. The transients are processed so that the transient energy is mostly contained on a single synthesis frame, effectively suppressing the transient smearing artifact which is typical for the phase vocoder based TSM.

Detection
To detect transients, the overall transientness of each analysis frame is estimated, and denoted as frame transientness:

$$r_t[m] = \frac{1}{N-1} \sum_{k=1}^{N-1} R_t[m,k]. \tag{61}$$

The analysis frames which are centered on a transient component appear as local maxima in the frame transientness. Transients need to be detected as soon as the analysis window slides over them in order to prevent the smearing of transient onsets. To this end, the time derivative of frame transientness is used:

$$\frac{d}{dm} r_t[m] \approx \frac{1}{H_a}(r_t[m] - r_t[m-1]), \tag{62}$$

where we approximate the time derivative with the backward difference method. As the analysis window slides over a transient, there is an abrupt increase in the frame transientness. These instants appear as local maxima in the time derivative of the frame transientness. Local maxima in the time derivative of the frame transientness that exceed a given threshold, are used for transient detection.

Figure 27 illustrates the proposed transient detection method using the same audio excerpt as above, containing piano, percussion, and double bass. The transients appear as local maxima in the frame transientness signal in Figure 27a. Transient onsets are detected from the time derivative of the frame transientness, from the local maxima which exceed the given threshold (the red dashed line in Figure 27b. The detected transient onsets are marked with orange crosses. After an onset is detected, the analysis frame which is centered on the transient is detected from the subsequent local maxima in the frame transientness. The detected analysis frames centered on a transient are marked with purple circles in Figure 27a.

Preservation
To prevent transient smearing, it is necessary to concentrate the transient energy in time. A single transient contributes energy to multiple analysis frames, because the frames are overlapping. During the synthesis, the phases of the STFT are modified, and the frames are relocated in time, which results in smearing of the transient energy.

To remove this effect, transients are detected as the analysis window slides over them. When a transient onset has been detected using the method described above,

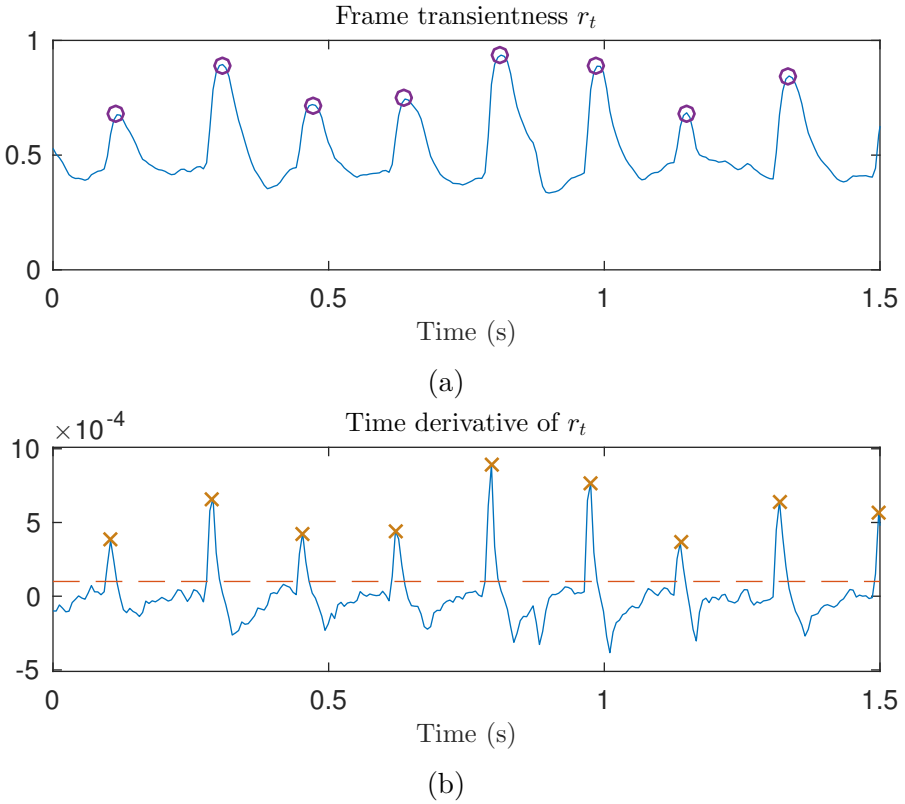


Figure 27: Illustration of the proposed transient detection. (**a**) Frame transientness. Locations of the detected transients are marked with purple circles. (**b**) Time derivative of the frame transientness. Detected transient onsets are marked with orange crosses. The red dashed line shows the transient detection threshold.

the energy in the STFT bins is suppressed according to their estimated transientness:

$$|Y[m,k]| = (1 - R_t[m,k])|X[m,k]|. \tag{63}$$

This gain is only applied to bins whose estimated transientness is larger than 0.5. Similar to [19], the bins to which this gain has been applied are kept in a non-contracting set of transient bins K_t. When it is detected that the analysis window is centered on a transient, as explained above, a phase reset is performed on the transient bins. That is, the original analysis phases are kept during synthesis for the transient bins. Subsequently, as the analysis window slides over the transient, the same gain reduction is applied for the transient bins as during the onset of the transient (63). The bins are retained in the set of transient bins until their transientness decays to a value smaller than 0.5, or until the analysis frame slides completely away from the detected transient center. Finally, since the synthesis frames before and after the center of the transient do not contribute to the transient's energy, the magnitudes of the transient bins are compensated by

$$|Y[m_t,k_t]| = \frac{\sum_{m\in\mathbb{Z}} w^2[(m_t-m)H_s]}{w^2[0]} \frac{\sum_{k\in K_t} R_t[m_t,k]}{|K_t|} |X[m_t,k_t]|, \tag{64}$$

where m_t is the transient frame index, $|K_t|$ denotes the number of elements in the set K_t, and $k_t \in K_t$, which is the defined set of transient bins.

This method aims to prevent the smearing of both the transient onsets and offsets during TSM. In effect, the transients are separated from the input audio, and relocated in time according to the TSM factor. However, in contrast to methods where transients are explicitly separated from the input audio [50, 49, 62, 41], the proposed method is more likely to keep transients perceptually intact with other components of the sound. Since the transients are kept in the same STFT representation, phase modifications in subsequent frames are dependent on the phases of the transient bins. This suggests that transients related to the onsets of harmonic sounds, such as the pluck of a note while strumming a guitar, should blend smoothly with the following tonal component of the sound. Furthermore, the soft manner in which the amplitudes of the transient bins are attenuated during onsets and offsets should prevent strong artifacts arising from errors in the transient detection.

Figure 28 shows an example of a transient processed with the proposed method. The original audio shown in Figure 28a consists of a solo violin overlaid with a castanet click. Figure 28b shows the time-scale modified sample with TSM factor $\alpha = 1.5$, using the standard phase vocoder. In the modified sample, the energy of the castanet click is spread over time. This demonstrates the well known transient smearing artifact of standard phase vocoder TSM. Figure 28c shows the time-scale modified sample using the proposed method. It can be seen that while the duration of the signal has changed, the castanet click in the modified audio resembles the one in the original, without any visible transient smearing.

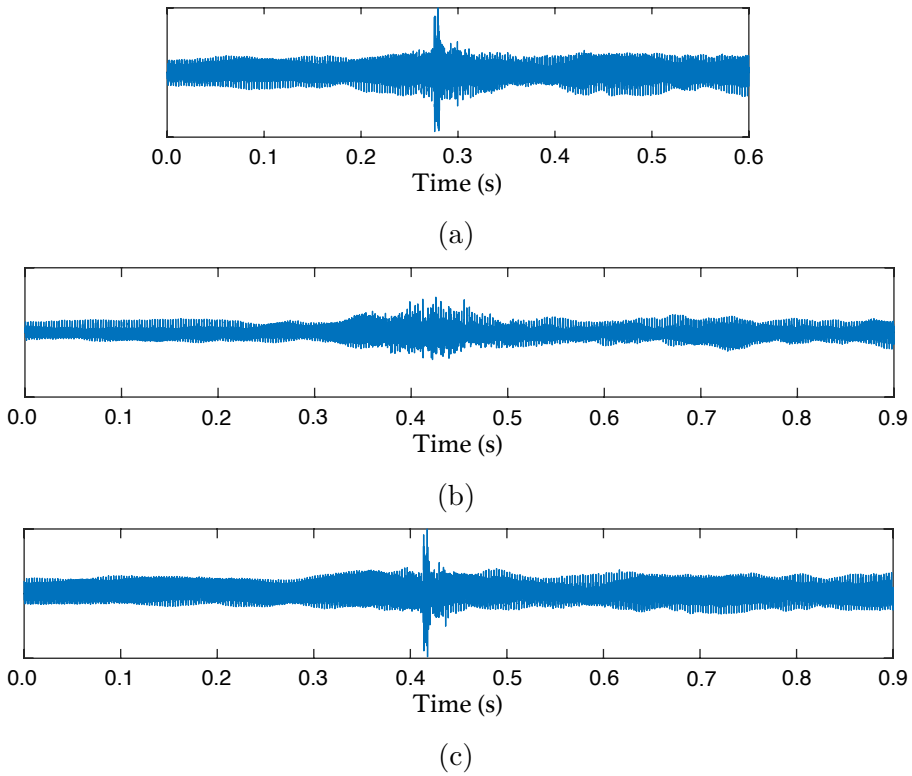


Figure 28: An example demonstrating the proposed transient preservation method. (**a**) shows the original audio, consisting of a solo violin overlaid with a castanet click. Also shown are the modified samples with TSM factor $\alpha = 1.5$, using (**b**) the standard phase vocoder, and (**c**) the proposed method.

6 Listening Test

To evaluate the quality of the proposed TSM technique, a listening test was conducted. The listening test was realized online using the Web Audio Evaluation Tool [63]. The test subjects were asked to use headphones. The test setup used was the same as in [41]. On each trial, the subjects were presented with the original audio sample and four modified samples processed with different TSM techniques. The subjects were asked to rate the quality of time-scale modified audio excerpts using a scale from 1 (bad) to 5 (excellent). A screenshot of the listening test environment is included in Appendix A.

All 11 subjects who participated in the test reported having a background in acoustics, and 10 of them had previous experience of participating in listening tests. None of the subjects reported hearing problems. The ages of the subjects ranged from 23 to 37, with a median age of 28. From the 11 subjects, ten were male and one was female.

In the evaluation of the proposed method, the following settings were used: the sample rate was 44.1 kHz, a Hann window of length $N = 4096$ was chosen for the STFT analysis and synthesis, the synthesis hop size was set to $H_s = 512$, and the number of frequency bins in the STFT was $K = N = 4096$. The length of the median filter in the frequency direction was 500 Hz, which corresponds to 46 bins. In the time direction, the length of the median filter was chosen to be 200 ms, but the number of frames it corresponds to depends on the analysis hop size, which is determined by the TSM factor according to (1). Finally, the transient detection threshold was set to $t_d = 10^{-4} = 0.00010$.

In addition to the proposed method (PROP), the following techniques were included: The standard phase vocoder (PV), using the same STFT analysis and synthesis settings as the proposed method; a recently published technique (HP) [41], which uses harmonic and percussive separation for transient preservation; the élastique algorithm (EL) [64], which is a state-of-the-art commercial tool for time and pitch-scale modification. The samples processed by these methods were obtained using the TSM toolbox [65].

Eight different audio excerpts (sampled at 44.1 kHz) and two different stretching factors $\alpha = 1.5$ and $\alpha = 2.0$ were tested, for the four techniques. This resulted in a total of 64 samples rated by each subject. The audio excerpts are described in Table 1. The lengths of the original audio excerpts ranged from 3 to 10 seconds. To estimate the quality of the techniques, mean opinion scores (MOS) were computed for all samples from the ratings given by the subjects. The results are shown in Table 2. A bar diagram of the mean opinion scores is also shown in Figure 29.

As expected, the standard PV performed worse than all the other tested methods. For the *CastViolin* sample, the proposed method (PROP) performed better than the other methods, with both TSM factors. This suggests that the proposed method preserves the quality of the transients in the modified signals better than the other methods. The proposed method also scored best with the *Jazz* excerpt. In addition to the well-preserved transients, the results are likely to be explained by the naturalness of the singing voice in the modified signals. This can be attributed to the proposed

Table 1: List of audio excerpts used in the subjective listening test.

Name	**Description**
CastViolin	Solo violin and castanets, from [65]
Classical	Excerpt from *Bólero*, performed by the *London Symphony Orchestra*
JJCale	Excerpt from *Cocaine*, performed by *J.J. Cale*
DrumSolo	Solo performed on a drum set, from [65]
Eddie	Excerpt from *Early in the Morning*, performed by *Eddie Rabbit*
Jazz	Excerpt from *I Can See Clearly*, performed by the *Holly Cole Trio*
Techno	Excerpt from *Return to Balojax*, performed by *Deviant Species and Scorb*
Vocals	Excerpt from *Tom's Diner*, performed by *Suzanne Vega*

Table 2: Mean opinion scores (MOS) for the audio samples.

	$\alpha = 1.5$				$\alpha = 2.0$			
	PV	HP	EL	PROP	PV	HP	EL	PROP
CastViolin	1.8	3.8	3.6	**4.1**	1.4	3.6	3.3	**4.1**
Classical	2.3	3.5	**3.7**	3.3	1.6	3.0	**3.7**	2.8
JJCale	2.7	2.5	**3.4**	2.9	1.2	2.5	3.1	**3.2**
DrumSolo	1.5	**3.5**	3.2	2.3	1.7	2.4	**2.5**	1.8
Eddie	1.9	3.1	**4.2**	3.2	1.2	2.2	**3.6**	3.1
Jazz	1.9	**3.6**	3.4	**3.6**	1.5	3.3	2.7	**3.7**
Techno	1.3	2.7	3.3	**4.1**	1.6	2.5	**3.1**	2.7
Vocals	1.7	**3.5**	2.9	3.4	1.5	**3.3**	2.7	3.1
Mean	1.9	3.3	**3.5**	3.4	1.5	2.9	**3.1**	**3.1**

phase propagation, which allows simultaneous preservation of the tonal and noisy qualities of the singing voice. This is also reflected in the results of the *Vocals* excerpt, where the proposed method also performs well, scoring slightly lower than HP, however. For the *Techno* sample, the proposed method scored significantly higher than the other methods with the TSM factor $\alpha = 1.5$. For TSM factor $\alpha = 2.0$, however, the proposed method scored lower than EL. The proposed method also scored highest for the *JJCale* sample with TSM factor $\alpha = 2.0$.

The proposed method performed poorer on the excerpts *DrumSolo* and *Classical*. Both of these samples contained fast sequences of transients. It is likely that the poorer performance is due to the individual transients not being resolved during the analysis, because of the relatively long analysis window used. Also on the excerpt *Eddie*, EL scored higher than the proposed method.

The preferences of subjects over the tested TSM methods seem to depend significantly on the signal being processed. Overall, the mean values computed from all the samples suggest that the proposed method yields a slightly better quality than HP with the large TSM factor $\alpha = 2.0$, and practically the same quality as EL. The processed audio excerpts are available online at http://research.spa.aalto.fi/publications/papers/applsci-ats/.

The proposed method introduces some additional computational complexity

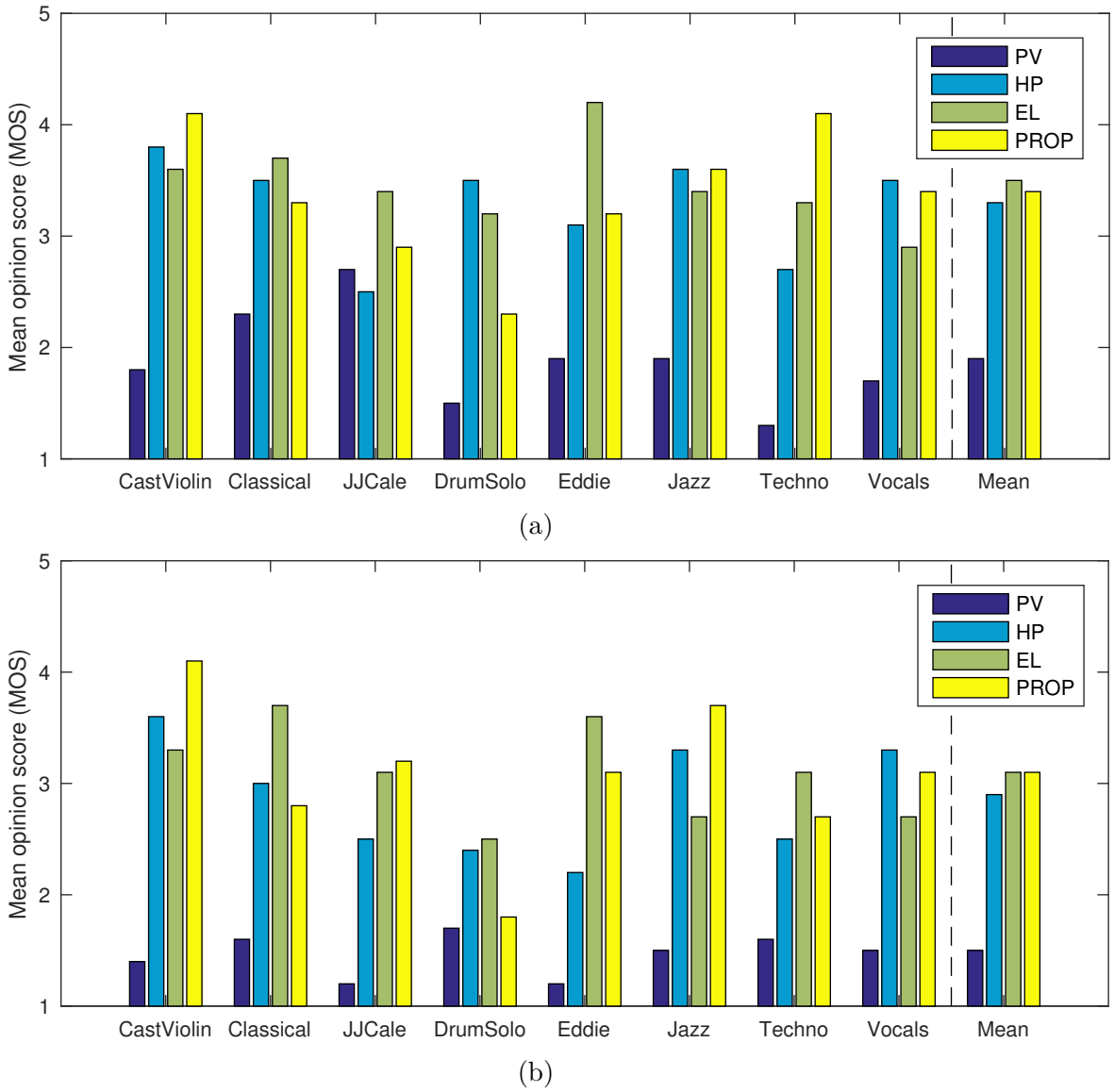


Figure 29: Mean opinion scores (MOS) for eight audio samples using four TSM methods for (**a**) medium ($\alpha = 1.5$) and (**b**) large ($\alpha = 2.0$) TSM factors. The rightmost bars show the average score for all eight samples. (PV = Phase Vocoder; HP = Harmonic-Percussive Separation [41]; EL = Élastique [64]; PROP = Proposed—this work).

when compared to the standard phase-locked vocoder. In the analysis stage, the fuzzy classification of the spectral bins requires median filtering of the magnitude of the analysis STFT. The number of samples in each median filtering operation depends on the analysis hop size and the number of frequency bins in each short time spectra. In the modification stage, additional complexity arises from drawing pseudo-random values for the phase randomization. Furthermore, computing the phase randomization factor, as in Equation (60), requires the evaluation of two hyperbolic tangent functions for each point in the STFT. Since the argument for the

second hyperbolic tangent depends only on the TSM factor, its value needs to be updated only when the TSM factor is changed. Finally, due to the way the values are used, a lookup table approximation can be used for evaluating the hyperbolic tangents without significantly affecting the quality of the modification.

7 Conclusions

In this thesis, a novel TSM technique was developed. The technique is based on the new concept of fuzzy classification of spectral bins. In the proposed classification scheme, each bin in the STFT representation of the input signal is assigned to three classes: tonalness, noisiness and transientness. The bins are allowed to belong to all of the classes simultaneously, with a certain degree of membership for each class which is estimated from the STFT representation. The information from the classification stage is used to guide the magnitude and phase modifications that are applied to the STFT during TSM, such that the subjective quality of the tonal, noise, and transient components is preserved.

By means of a listening test, the proposed method was compared to three TSM methods: the standard phase vocoder (PV), a state-of-the-art academic method (HP), and a commercial software (EL). The proposed method achieved higher scores than the standard phase vocoder with all processed audio excerpts. Overall, the proposed method performed slightly better than HP, and scored similarly as EL. For all methods, the quality of the transformation seemed to be highly dependent on the type of signal being processed. The proposed method performed best with samples such as a solo violin overlaid with a castanet, a jazz recording with vocals and with a techno song. The method performed worse when applied to a classical music recording, and to a drum solo.

While the method can preserve the quality of highly time-localized transients, such as the castanet click even using a large analysis window, it still suffers to some extent on the fixed time and frequency resolution of the STFT. With an input signal containing a fast sequence of transients, the individual transients can not be resolved in the STFT analysis if the time difference between consecutive transients is less than the length of the analysis window. Thus, applying the proposed techniques on a multiresolution time-frequency transformation is of future research interest. Finally, while this thesis only considered TSM, the proposed method for fuzzy classification of spectral bins could be useful in various audio signal processing problems, where information about the nature of the input signal is needed.

References

[1] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech communication*, vol. 16, no. 2, pp. 175–205, 1995.

[2] D. Barry, D. Dorran, and E. Coyle, "Time and pitch scale modification: A real-time framework and tutorial," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 103–110, Espoo, Finland, 1-4 September 2008.

[3] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, no. 2, p. 57, 2016.

[4] A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and G. Cohen, "Using audio time scale modification for video browsing," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS)*, Maui, HI, USA, 4-7 January 2000.

[5] D. Cliff, "Hang the DJ: Automatic sequencing and seamless mixing of dance-music tracks," *Hewlett-Packard Laboratories Technical Report*, vol. 104, 2000.

[6] H. Ishizaki, K. Hoashi, and Y. Takishima, "Full-automatic DJ mixing system with optimal tempo adjustment based on measurement function of user discomfort.," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 26–30 October 2009.

[7] O. Donnellan, E. Jung, and E. Coyle, "Speech-adaptive time-scale modification for computer assisted language-learning," in *Proceedings of the Third IEEE International Conference on Advanced Learning Technologies*, pp. 165–169, Athens, Greece, 9–11 July 2003.

[8] P. Dutilleux, G. De Poli, A. von dem Knesebeck, and U. Zölzer, "Time-segment processing (chapter 6)," in *DAFX: Digital Audio Effects, Second Edition* (U. Zölzer, ed.), pp. 185–217, Chichester, UK: Wiley, 2011.

[9] A. Moinet, T. Dutoit, and P. Latour, "Audio time-scaling for slow motion sports videos," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 314–320, Maynooth, Ireland, 2–5 September 2013.

[10] A. Haghparast, H. Penttinen, and V. Välimäki, "Real-time pitch-shifting of musical signals by a time-varying factor using normalized filtered correlation time-scale modification (NFC-TSM)," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 7–13, Bordeaux, France, 10–15 September 2007.

[11] J. Santacruz, L. Tardón, I. Barbancho, and A. Barbancho, "Spectral envelope transformation in singing voice for advanced pitch shifting," *Applied Sciences*, vol. 6, p. 368, 2016.

[12] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, pp. 493–496, Tampa, FL, USA, 26–29 April 1985.

[13] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 554–557, Minneapolis, MN, USA, 27–30 April 1993.

[14] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[15] M. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 374–390, 1981.

[16] M. Puckette, "Phase-locked vocoder," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 222–225, New Paltz, NY, USA, 15–18 October 1995.

[17] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio processing*, vol. 7, no. 3, pp. 323–332, 1999.

[18] J. L. Flanagan and R. Golden, "Phase vocoder," *Bell Labs Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[19] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx)*, pp. 344–349, London, UK, 8–11 September 2003.

[20] F. F. Lee, "Time compression and expansion of speech by the sampling method," *J. Audio Eng. Soc*, vol. 20, no. 9, pp. 738–742, 1972.

[21] S. Lee, H. D. Kim, and H. S. Kim, "Variable time-scale modification of speech using transient information," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 1319–1322, Münich, Germany, 21–24 April 1997.

[22] P. H. Wong, O. C. Au, J. W. Wong, and W. H. Lau, "On improving the intelligibility of synchronized over-lap-and-add (SOLA) at low TSM factor," in *Proceedings of IEEE Region 10 Annual Conference on Speech and Image Technologies for Computing and Telecommunications (TENCON)*, vol. 2, pp. 487–490, Brisbane, Australia, 2-4 December 1997.

[23] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 11, pp. 2015–2018, Tokyo, Japan, 7–11 April 1986.

[24] C. Hamon, E. Mouline, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 238–241, Glasgow, UK, 23–26 May 1989.

[25] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[26] D. Arfib, F. Keiler, U. Zölzer, V. Verfaille, and J. Bonada, "Time-frequency processing (chapter 7)," in *DAFX: Digital Audio Effects, Second Edition* (U. Zölzer, ed.), pp. 219–278, Chichester, UK: Wiley, 2011.

[27] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[28] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 243–248, 1976.

[29] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.

[30] J. A. Moorer, "The use of the phase vocoder in computer music applications," *Journal of the Audio Engineering Society*, vol. 26, no. 1/2, pp. 42–45, 1978.

[31] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.

[32] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1983.

[33] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[34] J. O. Smith, *Spectral Audio Signal Processing*. Available online: http://ccrma.stanford.edu/~jos/sasp/, accessed 10th October 2017. online book, 2011 edition.

[35] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[36] J. Laroche and M. Dolson, "Phase-vocoder: About this phasiness business," in *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 19–22 October 1997.

[37] T. F. Quatieri, R. B. Dunn, and T. E. Hanna, "A subband approach to time-scale expansion of complex acoustic signals," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 6, pp. 515–519, 1995.

[38] J. Bonada, "Automatic technique in frequency domain for near-lossless time-scale modification of audio," in *Proceedings of the International Computer Music Conference (ICMC)*, pp. 396–399, Berlin, Germany, 27 August – 1 September 2000.

[39] C. Duxbury, M. Davies, and M. B. Sandler, "Improved time-scaling of musical audio using phase locking at transients," in *Proceedings of the Audio Engineering Society 112th Convention*, München, Germany, 10–13 May 2002.

[40] C. Duxbury, M. Davies, and M. Sandler, "Extraction of transient content in musical audio using multiresolution analysis techniques," in *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, pp. 1–4, Limerick, Ireland, 6–8 December 2001.

[41] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.

[42] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 217–220, Graz, Austria, 6–10 September 2010.

[43] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, 1992.

[44] A. Röbel, "A shape-invariant phase vocoder for speech transformation," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 298–305, Graz, Austria, 6–10 September 2010.

[45] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[46] T. Quatieri and R. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1449–1464, 1986.

[47] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[48] T. S. Verma and T. H. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3573–3576, Las Vegas, NV, USA, 30 March – 4 April 1998.

[49] T. S. Verma and T. H. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music Journal*, vol. 24, no. 2, pp. 47–59, 2000.

[50] T. S. Verma and T. H. Meng, "Time scale modification using a sines+transients+noise signal model," in *Proceedings of the Digital Audio Effects Workshop (DAFx)*, Barcelona, Spain, 19–21 November 1998.

[51] M. Liuni, A. Robel, E. Matusiak, M. Romito, and X. Rodet, "Automatic adaptation of the time-frequency resolution for sound analysis and re-synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 959–970, 2013.

[52] P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus, and G. Velasco, "Theory, implementation and applications of nonstationary Gabor frames," *Journal of computational and applied mathematics*, vol. 236, no. 6, pp. 1481–1496, 2011.

[53] M. Dörfler and E. Matusiak, "Nonstationary Gabor frames—existence and construction," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 12, no. 03, p. 1450032, 2014.

[54] R. G. Baraniuk, P. Flandrin, A. J. Janssen, and O. J. Michel, "Measuring time-frequency information content using the Rényi entropies," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1391–1409, 2001.

[55] E. S. Ottosen and M. Döerfler, "A phase vocoder based on nonstationary Gabor frames," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2199–2208, 2017.

[56] N. Juillerat and B. Hirsbrunner, "Audio time stretching with an adaptive multiresolution phase vocoder," in *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 5–9 March 2017.

[57] N. Juillerat, S. M. Arisona, and S. Schubiger-Banz, "Enhancing the quality of audio transformations using the multi-scale short-time Fourier transform," in *Proceedings of the 10th IASTED International Conference on Signal and Image Processing*, vol. 623, p. 054, Kailua-Kona, HI, USA, 18–20 August 2008.

[58] Z. Průša and N. Holighaus, "Phase vocoder done right," in *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 28 August – 2 September 2017.

[59] Z. Průša and P. L. Søndergaard, "Real-time spectrogram inversion using phase gradient heap integration," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 17–21, Brno, Czech Republic, 5–9 September 2016.

[60] E.-P. Damskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Applied Sciences*, vol. 7, no. 12, p. 1293, 2017.

[61] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 1–4, Lausanne, Switzerland, 25–29 August 2008.

[62] F. Nagel and A. Walther, "A novel transient handling scheme for time stretching algorithms," in *Proceedings of the Audio Engineering Society 127th Convention*, New York, NY, USA, 9–12 October 2009.

[63] N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, "Web audio evaluation tool: A browser-based listening test environment," in *Proceedings of the 12th Sound and Music Computing Conference*, pp. 147–152, Maynooth, Ireland, 26 July – 1 August 2015.

[64] Zplane Development, "Élastique time stretching & pitch shifting SDKs." Available online: http://www.zplane.de/index.php?page=description-elastique. (accessed 20th October 2017).

[65] J. Driedger and M. Müller, "TSM toolbox: MATLAB implementations of time-scale modification algorithms," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 249–256, Erlangen, Germany, 1–5 September 2014.

A Listening Test Environment

A screenshot of the listening test environment is shown in Figure A1. The listening test was created using the Web Audio Evaluation Tool [63]. The participants rated the samples using the vertical sliders, which were initially set to random positions on each page. The participants had to listen to all samples on each page in order to move onto the next one. However, they did not have to listen to the samples fully. Furthermore, the participants had to move all the sliders on each page before allowing to continue onto the next page.

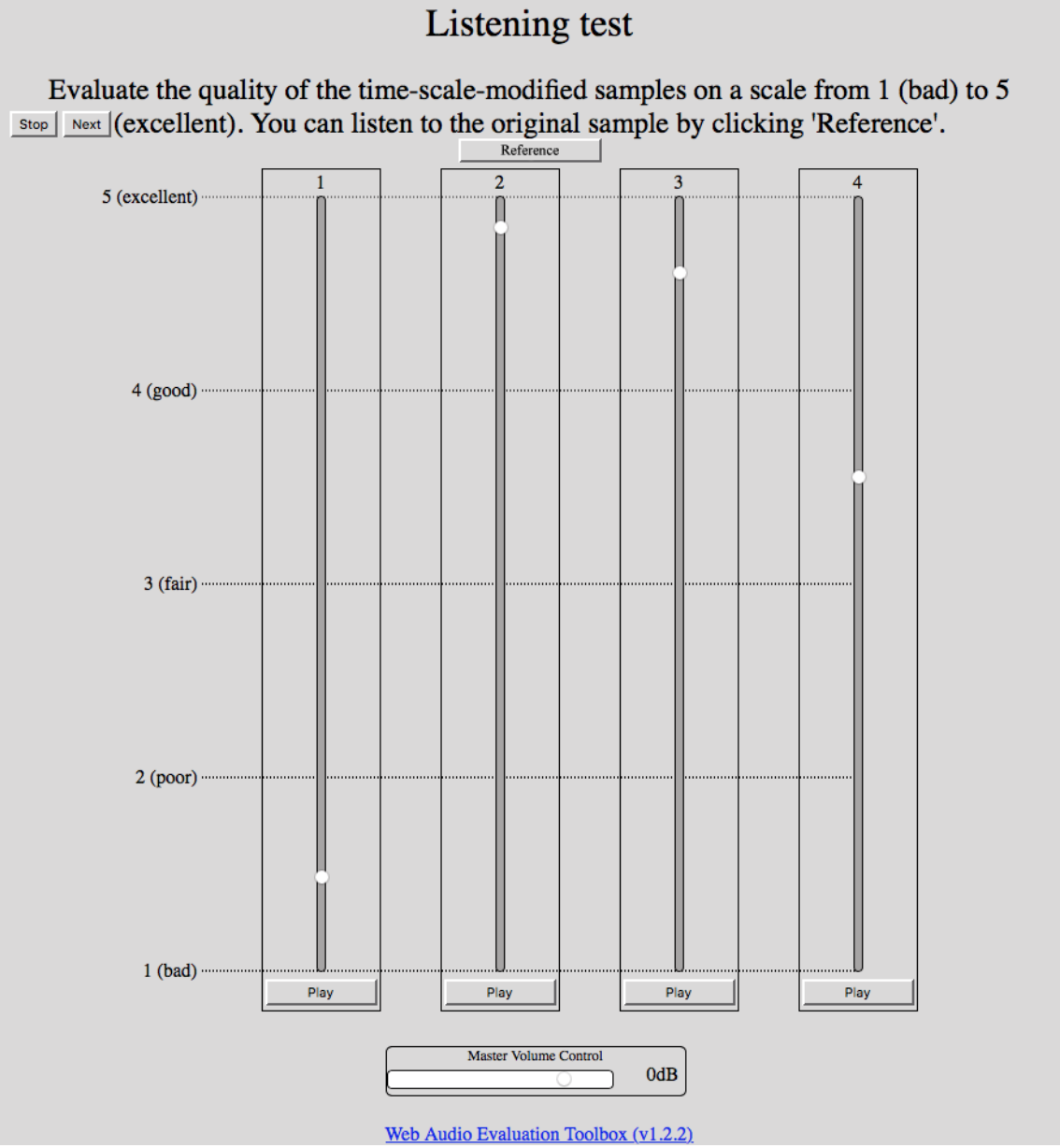


Figure A1: Screenshot of the listening test environment.