Improving Time-Scale Modification of Music Signals using Harmonic-Percussive Separation

Jonathan Driedger, Meinard Müller, Member, IEEE, and Sebastian Ewert,

Abstract—A major problem in time-scale modification (TSM) of music signals is that percussive transients are often perceptually degraded. To prevent this degradation, some TSM approaches try to explicitly identify transients in the input signal and to handle them in a special way. However, such approaches are problematic for two reasons. First, errors in the transient detection have an immediate influence on the final TSM result and, second, a perceptual transparent preservation of transients is by far not a trivial task. In this paper we present a TSM approach that handles transients implicitly by first separating the signal into a harmonic component as well as a percussive component. While the harmonic component is modified with a phase vocoder approach using a large frame size, the noise-like percussive component is modified with a simple time-domain overlap-add technique using a short frame size, which preserves the transients to a high degree without any explicit transient detection.

Index Terms—Time-scale modification, harmonic-percussive separation, transient preservation, phase vocoder, overlap-add

I. INTRODUCTION

The manipulation of the time-scale of a music audio signals without altering its pitch is a common task in music production or music remixing. Basic TSM algorithms like the phase vocoder [1] or WSOLA [2] succeed in solving this task to a certain degree, but not without introducing noticeable artifacts into the modified signal. Especially for transients as created by percussive instruments at note onsets, these algorithms usually fail to preserve the characteristics of the original sound. Therefore, several TSM algorithms have been proposed to overcome this problem by first explicitly identifying transients in the input signal and then giving them special treatment during the TSM process [3], [4], [5]. For example in [5], the identified transients are first removed from the signal together with a small tolerance region around each transient. The gaps in the remaining waveform are then closed by using linear prediction techniques and the resulting signal is modified using the phase vocoder. Afterwards, the unmodified regions containing the transients are temporally relocated according to the new time-scale and reinserted into the signal. Although this procedure is in principle capable of preserving correctly identified transients perfectly, errors in the transient detection have an immediate impact on the TSM result. Furthermore, there often occur artifacts at the boundaries of the reinserted segments.

In this paper, we propose a simple, yet effective method that avoids an explicit detection of transients, but is still capable of preserving them to a high degree. To this end, we first decompose the audio signal into a harmonic and a percussive component, using a recent computationally efficient algorithm [6]. One main observation is that the percussive component contains, besides other noise-like sounds, all the transients. The two components are then modified with two different TSM algorithms. The harmonic component is processed with a phase vocoder technique working with large frames, a technique that yields good TSM results as long as there are no sudden events in the audio signal. As for the percussive component, we found out that applying a simple time-domain overlap-add method (OLA) yields surprisingly good TSM results when using a very small frame size. Although OLA is considered a poor TSM technique in particular for harmonic sounds, it is capable of preserving the sound of transients without any explicit transient detection. Finally, the two modified components are superimposed to form the output of our procedure, see Figure 1. Even though OLA may introduce artifacts in case some harmonic sounds remain in the percussive component, they are often perceptually masked by the TSM result of the harmonic component and vice versa. To evaluate our proposed method, we performed a listening experiment. The results suggest that our relatively simple approach is almost always superior to the native phase vocoder or WSOLA and can also compete with a commercial state-of-the-art TSM
algorithm.

The remainder of this paper is structured as follows. In Section II we review the used harmonic-percussive separation procedure. Section III is devoted to TSM with a brief explanation of the applied algorithms as well as a discussion of parameter settings for our proposed combined method. The results of the conducted listening experiment as well as pointers to demo material are given in Section IV. Finally, in Section V, we wrap up this paper with conclusions and future work.

II. HARMONIC-PERCUSSIVE SEPARATION

Musical sounds can broadly be classified to be of harmonic or percussive nature. In a spectral representation, harmonic sounds manifest themselves as horizontal structures (in time direction) while percussive sounds have a vertical structure (in frequency direction). The goal of harmonic-percussive separation is to split an input audio signal $x$ into its harmonic component $x_h$ and its percussive component $x_p$ such that $x = x_h + x_p$. While there have been several approaches to solve this task (see for example [7], [8]), the approach of [6] is particularly efficient and elegant. We therefore use it in our proposed TSM algorithm and describe it briefly in the following.

First, a spectrogram $X$ of the signal $x$ is computed by applying the short-time Fourier transform (STFT)

$$X(t,k) = \sum_{n=0}^{N-1} w(n)x(n+tH)\exp(-2\pi ink/N) \quad (1)$$

with $t \in [0 : T - 1]$ and $k \in [0 : K]$. $T$ is the number of frames, $K = N/2$ is the frequency index corresponding to the Nyquist frequency, $N$ is the frame size and length of the discrete Fourier transform, $w$ is a window function and $H$ is the hopsize. Looking at one frequency band in the magnitude spectrogram $Y = |X|$ (one row of $Y$), harmonic components stay rather constant, while percussive structures yield peaks in the sequence. Contrary, in one frame (one column of $Y$), noise-like percussive components are equally distributed over the whole sequence, while the harmonic components stand out. By applying a median filter to $Y$ once in horizontal and once in vertical direction, we get a harmonically enhanced magnitude spectrogram $\hat{Y}_h$ and a percussively enhanced magnitude spectrogram $\hat{Y}_p$:

$$\hat{Y}_h(t,k) := \text{median}(Y(t - \ell,k), \ldots, Y(t + \ell,k)) \quad (2)$$

$$\hat{Y}_p(t,k) := \text{median}(Y(t,k - \ell), \ldots, Y(t,k + \ell)) \quad (3)$$

for some $\ell \in \mathbb{N}$ such that $2\ell + 1$ is the length of the median filter. From these, we can construct binary masks $M_h := (\hat{Y}_h \geq \hat{Y}_p)$ and $M_p := (\hat{Y}_p > \hat{Y}_h)$ where the operators $\geq$ and $>$ are applied point-wise and have the range $\{0,1\}$. Applying these masks to the original spectrogram $X$ yields the spectrograms for the harmonic and the percussive component of the signal $X_h := (X \odot M_h)$ and $X_p := (X \odot M_p)$, where the operator $\odot$ denotes point-wise multiplication. These spectrograms can then be brought back to the time-domain by applying an “inverse” short-time Fourier transform, see [9]. This yields the desired signals $x_h$ and $x_p$. For an overview of the procedure, see Figure 2.

III. TIME-SCALE MODIFICATION

Time-scale modification is the task of manipulating an audio signal such that it sounds as if its content was performed at a different tempo. TSM algorithms usually achieve this by segmenting an input signal $x$ into $T$ short overlapping frames $x_t$ of length $N$, spaced apart by a fixed analysis hopsize $H_a$.

$$x_t(m) = \begin{cases} x(m + tH_a) & \text{for } m \in [-N/2 : N/2 - 1] \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Depending on the TSM algorithm, the frames are suitably modified to compensate for phase discontinuities, yielding frames $y_t$. These are then multiplied with a window function $w$, added up with a synthesis hopsize $H_s$, and normalized to form the output signal $y$. Formally, $y$ is defined by

$$y(m) = \frac{\sum_{t=0}^{T-1} w(m - tH_s) y_t(m - tH_a)}{\sum_{t=0}^{T-1} w(m - tH_s)} \quad (5)$$

The signal $y$ is a time-scale modified version of $x$, altered in length by a factor of $\alpha = H_s/H_a$. A method to compute the frames $y_t$ from $x_t$ is the phase vocoder [1]. It manipulates the phases of the spectrogram $X$ of $x$ to ensure horizontal phase coherence for all frequency bands with the synthesis hopsize $H_s$. We omit the details of the phase vocoder at this point and instead refer to [10], where also further refinements and modifications of the procedure are introduced. By design, the phase vocoder is well suited for modifying signals of harmonic nature. Transients on the other hand are often smeared by the phase vocoder due to the loss of vertical phase coherence.
during the phase adaption process. Since the phase vocoder relies on a high frequency resolution of the spectrogram $X$, and therefore a large frame size $N$, this smearing effect cannot be reduced by simply reducing the frame size.

Contrary to the phase vocoder, the basic overlap-add (OLA) TSM, which can be defined by setting $y_t = x_t$ in Equation 5, is usually considered a rather poor TSM approach. Indeed, for general signals it produces strong artifacts, in particular for harmonic sounds. The reason for this is that OLA is not capable of preserving periodic structures in a signal and introduces phase jumps. However, we have made the interesting observation that OLA preserves the sound of noise-like signals and transients quite well, especially when using a small frame size. Indeed, here the introduced phase jumps do not influence the rather chaotic nature of phases and therefore have no impact of perceptual relevance. Furthermore, when using a very small frame size, there are nearly no stuttering artifacts perceivable at transients (see for example Figure 3b). Such artifacts usually occur in TSM results obtained by time-domain TSM algorithms like WSOLA, which rely on larger frame sizes (see Figure 3f). Although in the case of OLA the transients are slightly stretched in length as well, they are still perceptually appealing and crisp for reasonable $\alpha$ (we experimented with $\alpha \in [0.5, 3]$). The conclusion is that OLA can be used to modify noise-like signals, while preserving transients without any explicit transient detection step.

In our proposed method, the idea is to split up the input signal into a harmonic and a percussive component and then to apply two specialized algorithms to the respective components: the phase vocoder using a large frame size to the harmonic component $x_h$ and OLA using a small frame size to the percussive component $x_p$. The superimposition of the two modified components forms the final TSM result of our procedure.

Concerning parameters settings, there are a few points to notice. For the harmonic-percussive separation, the most crucial parameter to influence the separation result is the frame size $N$ of the STFT computation. The larger $N$, the larger the portion of the signal that is assigned to the harmonic component $x_h$. This is the case since a transient is an event of very short temporal duration and its influence on the structure of the magnitude spectrogram $Y$ therefore decreases when $N$ becomes larger. In our implementation, we set $N$ to 1024 samples while we use a sampling rate of $22050$ Hz for the original music signals. The length of the median filter $2\ell + 1$ on the other hand does not influence the separation too much as long as no extreme values are chosen. We set $\ell = 5$ in our experiments.

For the phase vocoder, it is important to use large frame sizes $N$ to get a good frequency resolution in the STFT computation. We therefore use a frame size of 4096 samples. Furthermore, we implemented the identity phase locking strategy presented in [10] to increase the overall quality of the technique. Finally, an important observation is that applying the window function $w$ to the frames $y_t$ usually leads to a loss in energy of the output signal. We therefore rescale the amplitude of frames $y_t$ to compensate for that loss prior to the synthesis of $y$ described by Equation 5. Not doing this would result in a damped harmonic component in the output of the algorithm. Like mentioned above, for OLA it is crucial to use a very small frame size $N$ to ensure the accurate temporal location and sound preservation of the percussive signal elements. In our implementation we set $N$ to 256 samples.

### IV. Evaluation

To evaluate our proposed method, we conducted an online listening experiment where we compared five TSM algorithms: our proposed method based on harmonic-percussive separation (HP), the commercial elastique algorithm [11] (EL) which is included in a wide range of music software nowadays and can be considered one of the important state-of-the-art algorithms, the phase vocoder with identity phase locking [10] (PV), the phase vocoder with transient preservation by Nagel and Walther as described in [5] (NW), and WSOLA [2] (WS). With each of these algorithms the ten short audio excerpts listed in Table I were stretched with constant stretching factors $\alpha = 1.2$ and $\alpha = 1.8$ resulting in 100 test items in total. In the experiment the test listeners were sequentially presented with groups of five stretched items of one excerpt together with the original excerpt and asked to rate the five TSM results on the five point mean opinion score scale from 1 (bad) to 5 (excellent). Overall 22 people, most of them with a background in audio signal processing, participated in the experiment. The results are given in Table II.

One can see that the two algorithms HP and EL performed significantly better than the remaining three algorithms. For highly percussive excerpts like Bongo, CastanetsViolin and DrumSolo HP even outperformed all other tested TSM procedures. These good results can mainly be devoted to the well-preserved transients in the modified signals. Looking for example at the TSM results of all five algorithms for the CastanetsViolin excerpt in Figure 3 this becomes obvious. For EL, PV and WS (Figure 3c, d and f) one can observe stuttering or smearing artifacts at transient positions. NW (Figure 3e) is actually able to preserve the transients perfectly.

1 All sound files used in the experiment as well as a link to the online experiment itself can be found at http://www.audiolabs-erlangen.de/resources/2014-SPL-HPTSM/.

<table>
<thead>
<tr>
<th>Item name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bongo</td>
<td>Regular beat played on bongo.</td>
</tr>
<tr>
<td>CastanetsViolin</td>
<td>Solo violin overlayed with castanets.</td>
</tr>
<tr>
<td>DrumSolo</td>
<td>A solo performed on a drum set.</td>
</tr>
<tr>
<td>Glockenspiel</td>
<td>Monophonic melody played on a glockenspiel.</td>
</tr>
<tr>
<td>Jazz</td>
<td>Synthetic polyphonic sound mixture of a trumpet, a piano, a bass and drums.</td>
</tr>
<tr>
<td>Pop</td>
<td>Synthetic polyphonic sound mixture of several synthesizers, a guitar and drums.</td>
</tr>
<tr>
<td>SingingVoice</td>
<td>Solo male singing voice.</td>
</tr>
<tr>
<td>Stepdad</td>
<td>Excerpt from My Leather, My Fur, My Nails by the band Stepdad.</td>
</tr>
<tr>
<td>SynthMono</td>
<td>Monophonic synthesizer with a very noisy and distorted sound.</td>
</tr>
<tr>
<td>SynthPoly</td>
<td>Sound mixture of several polyphonic synthesizers.</td>
</tr>
</tbody>
</table>

**TABLE I**

LIST OF AUDIO EXCERPTS.
but the tolerance region around the actual transients, which are copied from the original and reinserted into the modified signal, make the TSM result sound very unnatural. This shows that even when transients are identified correctly, the re-insertion into the modified signal is another problematic step causing unnatural transitions and also explains the bad performance of NW in the listening experiment. In contrast, re-insertion into the modified signal is another problematic signal, make the TSM result sound very unnatural. This are copied from the original and reinserted into the modified but the tolerance region around the actual transients, which are neither of harmonic nor of percussive nature.

Table II

<table>
<thead>
<tr>
<th>Sample</th>
<th>HP</th>
<th>EL</th>
<th>PK</th>
<th>NW</th>
<th>WS</th>
<th>HP</th>
<th>EL</th>
<th>PK</th>
<th>NW</th>
<th>WS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bongo</td>
<td>4.73</td>
<td>4.32</td>
<td>1.68</td>
<td>2.27</td>
<td>1.64</td>
<td>4.09</td>
<td>3.50</td>
<td>1.36</td>
<td>1.59</td>
<td>1.41</td>
</tr>
<tr>
<td>CastanetsViolin</td>
<td>4.41</td>
<td>4.36</td>
<td>1.95</td>
<td>1.68</td>
<td>2.77</td>
<td>3.91</td>
<td>3.68</td>
<td>1.64</td>
<td>1.64</td>
<td>2.27</td>
</tr>
<tr>
<td>DrumSolo</td>
<td>4.27</td>
<td>3.77</td>
<td>3.09</td>
<td>1.95</td>
<td>1.68</td>
<td>2.95</td>
<td>3.09</td>
<td>1.77</td>
<td>1.32</td>
<td>1.32</td>
</tr>
<tr>
<td>Glockenspiel</td>
<td>3.18</td>
<td>4.45</td>
<td>2.86</td>
<td>1.82</td>
<td>2.18</td>
<td>3.14</td>
<td>3.91</td>
<td>2.64</td>
<td>1.55</td>
<td>2.00</td>
</tr>
<tr>
<td>Jazz</td>
<td>4.55</td>
<td>4.09</td>
<td>3.77</td>
<td>2.00</td>
<td>2.14</td>
<td>3.68</td>
<td>3.27</td>
<td>2.23</td>
<td>1.86</td>
<td>1.23</td>
</tr>
<tr>
<td>Pop</td>
<td>4.23</td>
<td>4.00</td>
<td>3.73</td>
<td>1.68</td>
<td>1.73</td>
<td>3.59</td>
<td>3.27</td>
<td>2.41</td>
<td>2.14</td>
<td>1.68</td>
</tr>
<tr>
<td>SingingVoice</td>
<td>5.79</td>
<td>4.27</td>
<td>3.32</td>
<td>1.73</td>
<td>3.82</td>
<td>2.91</td>
<td>3.77</td>
<td>2.91</td>
<td>1.45</td>
<td>2.32</td>
</tr>
<tr>
<td>Stepdad</td>
<td>4.32</td>
<td>3.82</td>
<td>3.64</td>
<td>1.82</td>
<td>2.14</td>
<td>3.86</td>
<td>3.36</td>
<td>2.91</td>
<td>1.73</td>
<td>1.82</td>
</tr>
<tr>
<td>SynthMono</td>
<td>2.82</td>
<td>2.05</td>
<td>2.32</td>
<td>1.95</td>
<td>2.82</td>
<td>2.45</td>
<td>3.09</td>
<td>2.50</td>
<td>1.95</td>
<td>2.27</td>
</tr>
<tr>
<td>SynthPoly</td>
<td>3.41</td>
<td>4.14</td>
<td>2.55</td>
<td>2.68</td>
<td>3.50</td>
<td>3.09</td>
<td>3.91</td>
<td>2.05</td>
<td>2.91</td>
<td>2.64</td>
</tr>
</tbody>
</table>

Fig. 3. (a): Excerpt of the CastanetsViolin Item. (b)-(f): TSM results for a constant stretching factor of 1.8 of (b): HP (e); EL (d): PV (e); NW (f): WS.

V. CONCLUSION AND FUTURE WORK

In this paper we presented a novel combination of known algorithms to improve the quality of TSM results for music signals. By separating an input signal into a harmonic and a percussive component, which are then modified using different TSM algorithms, no explicit transient detection is necessary to preserve transients in the recombined output signal. The approach of applying harmonic-percussive separation prior the actual processing step is not only beneficial for TSM, but may also be applied in other scenarios such as bandwidth extension or in concealment algorithms. Furthermore, an interesting future research direction is to better handle signals that have dominant components that are neither of harmonic nor of percussive nature.

REFERENCES


