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 which typically contains the transients. 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*—Harmonic-percussive separation, overlap-add, phase vocoder, time-scale modification, transient preservation.**

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 for harmonic signals but *smears* sudden events like transients noticeably [7]. 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 Fig. 1. Even though OLA may introduce artifacts in case some harmonic sounds remain in the percussive component due to an imperfect decomposition, those artifacts 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

Manuscript received September 23, 2013; revised November 19, 2013; accepted November 25, 2013. Date of publication December 05, 2013; date of current version December 13, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Augusto Sarti.

J. Driedger and M. Müller are with the International Audio Laboratories Erlangen. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen (IIS), Erlangen, Germany.

S. Ewert is with Queen Mary University of London, London, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2013.2294023

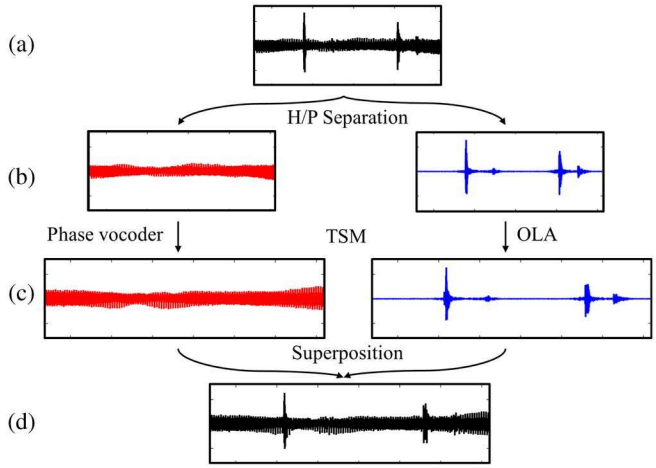

Fig. 1. Overview of the proposed TSM approach. (a): Input music recording x. (b): Separation in harmonic component x_h (left) and percussive component x_p (right). (c): TSM results for the harmonic component using the phase vocoder (left) and for the percussive component using OLA (right). (d): Superposition of the TSM results from (c).

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 [8], [9]), 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)

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

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 $\tilde{Y}_h$ and a percussively enhanced magnitude spectrogram $\tilde{Y}_p$:

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

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

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 $M_h := (\tilde{Y}_h \geq \tilde{Y}_p)$ and $M_p := (\tilde{Y}_p > \tilde{Y}_h)$ 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 [10]. This yields the desired signals x_h and x_p. For an overview of the procedure, see Fig. 2.

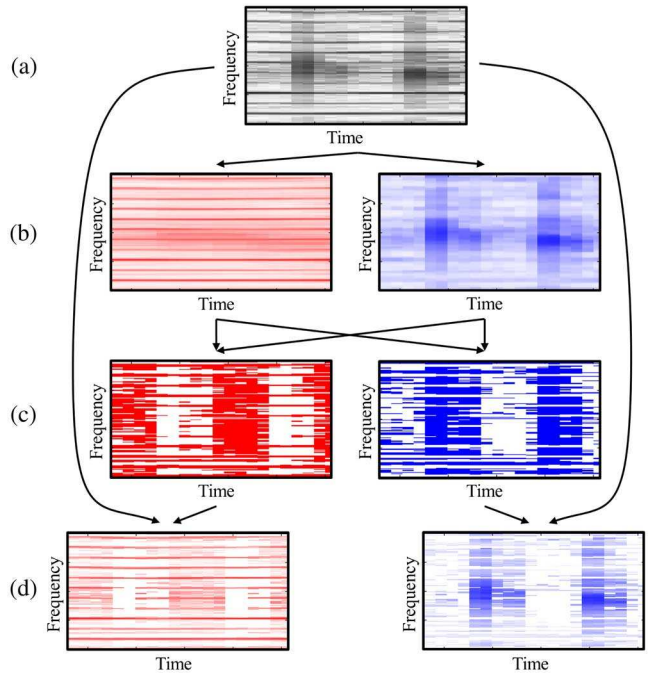

Fig. 2. Harmonic-percussive separation. (a): Spectrogram X. (b): Horizontally and vertically median-filtered magnitude spectrograms $\tilde{Y}_h$ (left) and $\tilde{Y}_p$ (right). (c): Binary masks M_h (left) and M_p (right) derived from $\tilde{Y}_h$ and $\tilde{Y}_p$. (d): Masked spectrograms X_h (left) and X_p (right).

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.

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

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

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

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 [7], 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

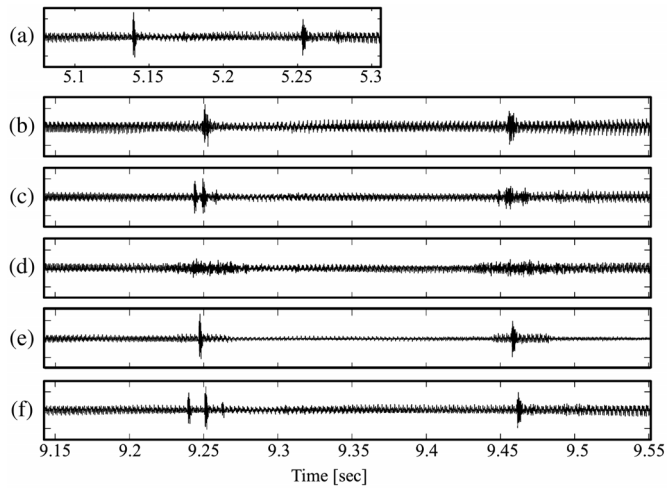

Fig. 3. (a): Excerpt of the *CastanetsViolin* Item. (b)-(f): TSM results for a constant stretching factor of 1.8 of (b): HP (c): EL (d): PV (e): NW (f): WS.

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 Fig. 3b). Such artifacts usually occur in TSM results obtained by time-domain TSM algorithms like WSOLA [2], which rely on larger frame sizes (see Fig. 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 α (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. 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. Our experiments have shown that setting $N = 1024$ and $\ell = 5$ using a sampling rate of 22050 Hz for the original music signals yields decompositions in which the percussive component contains the transients while showing only very few harmonic residual sounds. In further experiments it showed that for obtaining such results only the ratio between the sampling rate and the frame size is important. For example, we obtain perceptually similar decompositions by setting $N = 2048$ for signals sampled at 44100 Hz. Systematic experiments on choosing optimal values for N and ℓ are left as future work and may further improve the TSM results of our proposed procedure.

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 for signals sampled at 22050 Hz. Furthermore, we implemented the identity phase locking strategy presented in [7] 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 (all sources of the experiment can be found at [11]) where we compared five TSM algorithms: our proposed method based on harmonic-percussive separation (HP), the commercial *élastique* algorithm [12] (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 [7] (PV), the phase vocoder with transient preservation by Nagel and Walther as described in [5] (NW), and *WSOLA* [2] (WS). As evaluation dataset we used the ten short audio excerpts listed in Table I. The set was compiled to cover a wide range of aspects in music signals. Among the excerpts are purely percussive signals, monophonic as well as highly polyphonic signals with different instrumentations, simple synthetic sound mixtures as well as professionally mixed studio recordings. With each of above listed five algorithms the ten excerpts were stretched with a moderate stretching factor $\alpha = 1.2$ and a stronger stretching factor $\alpha = 1.8$ as they realistically occur in applications like audio remixing (TSM results for the more extreme stretching factors $\alpha = 0.5$ and $\alpha = 3$ can be found at [11]). This results 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

TABLE I
LIST OF AUDIO EXCERPTS

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

TABLE II
MEAN OPINION SCORES FOR ALL TEST ITEMS

	$\alpha = 1.2$					$\alpha = 1.8$				
	HP	EL	PV	NW	WS	HP	EL	PV	NW	WS
Bongo	**4.73**	4.32	1.68	2.27	1.64	**4.09**	3.50	1.36	1.59	1.41
CastanetsViolin	**4.41**	4.36	1.95	1.68	2.77	**3.91**	3.68	1.64	1.64	2.27
DrumSolo	**4.27**	3.77	3.09	1.95	1.68	2.95	**3.09**	1.77	1.32	1.32
Glockenspiel	3.18	**4.45**	2.86	1.82	2.18	3.14	**3.91**	2.64	1.55	2.00
Jazz	**4.55**	4.09	3.77	2.00	2.14	**3.68**	3.27	2.23	1.86	1.23
Pop	**4.23**	4.00	3.73	1.68	1.73	**3.59**	3.27	2.41	2.14	1.68
SingingVoice	2.59	**4.27**	3.32	1.73	3.82	2.91	**3.77**	2.91	1.45	2.32
Stepdad	**4.32**	3.82	3.64	1.82	2.14	**3.86**	3.36	2.91	1.73	1.82
SynthMono	2.82	**4.05**	2.32	1.95	2.82	2.45	**3.09**	2.50	1.95	2.27
SynthPoly	3.41	**4.14**	2.55	2.68	3.50	3.09	**3.91**	2.05	2.91	2.64
∅	3.85	4.13	2.89	1.96	2.44	3.37	3.49	2.24	1.81	1.90

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 Fig. 3 this becomes obvious. For EL, PV and WS (Fig. 3c, d and f) one can observe stuttering or smearing artifacts at transient positions. NW (Fig. 3e) is actually able to preserve the transients perfectly, 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, HP (Fig. 3b) succeeds in keeping the transients rather crisp while also preserving the harmonic component around the transients.

Also for complex sound mixtures with strong percussive components like *Jazz*, *Pop* and *Stepdad* HP scored best. This shows, that the harmonic-percussive separation actually performs well, also on very complex musical signals. Noticeable are the poorer performances of HP for the excerpts *SynthMono* and *SingingVoice*. For *SynthMono*, a closer investigation of HP's TSM results revealed that the modified harmonic components suffers from a well known artifact known as *phasiness*, which is a characteristic "coloration" of the sound introduced by the phase vocoder [7]. The strong presence of this artifact, which originates from the loss of phase coherence during the modification process, can be made responsible for the lower scores. Therefore, by reducing phasiness artifacts in the TSM results of the harmonic components one may further improve our proposed algorithm. For *SingingVoice*, it showed that for this excerpt the separation into a harmonic and a percussive component is problematic. Due to the spectral characteristics of human voice, which are neither of clearly harmonic nor clearly percussive nature, the portions of the spectrogram which are assigned to the percussive component still cover large amounts of harmonic content. Therefore, the assumption that the percussive component is mainly noise-like is violated and, when modifying this component with OLA, the introduced phase jump artifacts are too strong to be masked by the modified harmonic component. This also explains why the native phase vocoder PV scores better than HP for this excerpt. However, HP performed significantly better than PV on average and is comparable in quality to the proprietary and highly optimized EL on the tested dataset. Considering the simplicity of our method, this is quite remarkable. Our contribution shows that a robust and perceptually transparent preservation of transients is essential for a "good" TSM result. Furthermore, it shows that to achieve such a preservation, an explicit detection of transients is not necessary.

Concerning the running time, the harmonic-percussive separation can be implemented efficiently and the additional computation time is independent of the stretching factor α. For example, for stretching one minute of input audio material with stretching factors of $\alpha = 1.2$ and $\alpha = 1.8$, our MATLAB implementation of HP needed four seconds and five seconds, respectively.

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. Concerning future research, finding ways to reduce phasiness artifacts in the TSM results of the phase vocoder may further increase the quality of our proposed procedure. Another interesting research direction is to better handle signals that have dominant components that are neither of clearly harmonic nor of clearly percussive nature and concern more the "texture" of a sound [13]. Finally, 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.

REFERENCES

[1] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. l J.*, vol. 45, pp. 1493–1509, 1966.

[2] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE International Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Minneapolis, MN, USA, 1993.

[3] S. Grofit and Y. Lavner, "Time-scale modification of audio signals using enhanced WSOLA with management of transients," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 106–115, 2008.

[4] C. Duxbury, M. Davies, and M. Sandler, "Improved time-scaling of musical audio using phase locking at transients," in *Audio Engineering Soc. Conv. 112*, Apr. 2002.

[5] F. Nagel and A. Walther, "A novel transient handling scheme for time stretching algorithms," in *127th Audio Engineering Soc. Conv. 2009*, New York, NY, 2009, pp. 185–192.

[6] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Graz, Austria, 2010, pp. 246–253.

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

[8] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Philadelphia, PA, USA, 2008, pp. 139–144.

[9] C. Duxbury, M. Davies, and M. Sandler, "Separation of transient information in audio using multiresolution analysis techniques," in *Proc. COST G-6 Conference on Digital Audio Effects (DAFX-01)*, Limerick, Ireland, Dec. 2001.

[10] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[11] J. Driedger, M. Müller, and S. Ewert, Accompanying website: Improving time-scale modification of music signals using harmonic-percussive separation [Online]. Available: http://www.audiolabs-erlangen.de/resources/2014-SPL-HPTSM/

[12] zplane development, élastique time stretching & pitch shifting SDKs [Online]. Available: http://www.zplane.de/index.php?page=description-elastique Aug. 2013, Accessed

[13] N. Saint-Arnaud and K. Popat, "Computational auditory scene analysis," in *chapter Analysis and Synthesis of Sound Textures*. Hillsdale, NJ, USA: Lawrence Erlbaum, 1998, pp. 293–308.