AUDIO TIME STRETCHING WITH AN ADAPTIVE MULTIRESOLUTION PHASE VOCODER

Nicolas Juillerat, Béat Hirsbrunner

University of Fribourg, Switzerland

ABSTRACT

This paper presents a novel adaptive resolution technique for audio time stretching. The signal's transience is analyzed in both time and frequency, and controls the time-frequency resolution of the processing. Higher time resolution is used for transients and higher frequency resolution for steady sounds. To preserve phase coherence between components of different transience, only the magnitude is processed adaptively. This is done by first using a single resolution phase vocoder, and then by applying a multiresolution magnitude correction step.

The proposed technique uses a decomposition of the signal in both time and frequency according to several degrees of transience. Yet the actual time stretching transformation is entirely based on the phase vocoder.

Index Terms— Time stretching, transient, magnitude correction, multiresolution

1. INTRODUCTION

Time stretching of audio signals using a phase vocoder gives high quality results, but tends to smear the transients because of insufficient time resolution [1]. It is hence necessary to detect and process the transients differently.

The proposed method differs from previous phase vocoder techniques by combining three aspects.

First, transients are not only detected in time, but in both time and frequency. Hence a transient located in the high frequencies for example will not affect non-transient components that occur at the same time in the low frequencies.

Second, it does not use a binary transient / non transient separation, but splits the signal according to any number of degrees of transience. As a result, a weak transient mixed with a non-transient sound will be processed with an intermediate time-frequency resolution that is sufficient to perceptually preserve the transient, while minimally affecting the non-transient sound.

Third, to preserve phase coherence between components processed with different resolutions, only the magnitude is processed using adaptive resolutions. This is done by first processing both the phase and the magnitude using a single resolution phase vocoder, and then dynamically fixing the magnitude using multiple resolutions based on the analyzed transience.

The phase processing part of the presented algorithm reuses the phase vocoder with scaled phase locking without modification, as a black box. Hence it is not detailed here. The reader might refer to the work of Laroche and Dolson [1].

This paper is organized as follows: Chapter 2 discusses related work. Chapter 3 introduces the terminology and notations. The technique is detailed in Chapter 4, with specific problems and improvements in Chapters 5 and 6. Results are given in Chapter 7.

2. RELATED WORK

Transient processing with a phase vocoder is usually done by shifting the transients in time without transforming them [2], by resetting the phase at the transients [3], or by a combination of both [4].

Audio transformations based on the short-time Fourier transform can in general be performed with adaptive resolution by using different resolutions independently, and then using a "mixer of coefficients" that combines the results based on an analysis of the signal transience [5]. This general approach however is only applicable to transformations that do not change the phase, such as noise reduction or center channel extraction. Applying it to the phase vocoder would not preserve the phase coherence between the different resolutions, and introduces audible glitches around transients.

This phase problem can be solved using additional steps to lock the phase of the signals processed with different resolutions [6]. It can hence be applied to time stretching and results in high quality. However, that process requires such a huge overlapping factor that it is typically about 50 times slower than a plain phase vocoder, limiting its usefulness in practice. The technique proposed in this paper has similar properties in terms of quality but does not require high overlapping. The current implementation is only about 4 to 5 times slower than a plain phase vocoder.

A different approach is to process only the steady components with a phase vocoder, and to use a time-domain overlap-add scheme for the transients [7]. While this approach handles transients in both time and frequency, it cannot be accommodated to several degrees of transience.

It has also been proposed to process high frequencies with a smaller frequency resolution than low frequencies, without any analysis of the signal. The underlying idea is to better match human perception. This "pyramidal" resolution scheme is however not very effective in preventing transient smearing when used alone [4], [8].

While the technique introduced in this paper is based on a phase vocoder using short-time Fourier transforms, the problem of transient smearing with time stretching can also be handled using completely different approaches. Examples are high-level signal models based on sines, transients and noise [9], or techniques using nonstationary Gabor frames [10] or wavelet packets [11].

3. DEFINITIONS AND NOTATIONS

The phase vocoder as well as other parts of the technique presented in this paper use the short-time Fourier transform (STFT), which is defined by the length of the Fourier transform N, and the analysis and synthesis hop sizes R_a and R_s. The choice of N depends on the desired time-frequency resolution and on the sample rate f_s. The STFT resolution τ is given by the duration of the block that is Fourier transformed: $\tau = \frac{N}{f_s}$.

Given an audio signal $x[t]$ in the time domain, the STFT transforms it into a *spectrum* given by $X[s,k]$, where the coefficient s is the frame index and is related to time (in samples) by $t = sR_a$. $X[s,k]$ is a *bin*, a complex number that gives the phase and magnitude of the signal at frame index s and at frequency $k\frac{f_s}{N}$ in Hz. For a given frame index s, the list of values $X[s,k]$ for $k \in [0, \frac{N}{2}]$ is a *spectral frame*.

The *magnitude spectrum* is defined as $|X[k,s]|$.

4. DESCRIPTION OF THE TECHNIQUE

The key idea of the proposed technique consists of three points:

- process the phase with a single, high frequency resolution,
- process the magnitude with multiple, adaptive time-frequency resolutions based on the signal transience,
- combine the two to produce the final result.

The intuition behind these points is that phase is mostly important for steady sounds. Transients on the other hand are mostly defined by their fast magnitude variations and carry little useful phase information [7].

Note that processing phase alone is not practical; therefore it is instead proposed to process both phase and magnitude using a single resolution, and then to fix the resulting magnitude using that of a magnitude-only, multiresolution processing.

The detailed workflow of the proposed technique is illustrated in Figure 1. It consists of five different components:

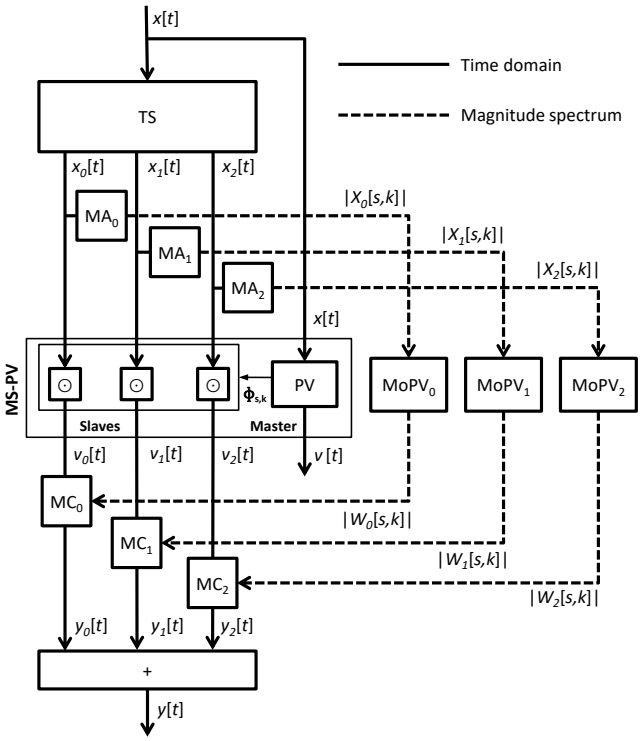

Figure 1: workflow of the proposed algorithm

- transience splitting (TS),
- magnitude analysis (MA),
- master-slave phase vocoder (MS-PV),
- magnitude-only phase vocoder (MoPV),
- magnitude correction (MC).

These components are detailed in the next subsections.

4.1. Transience splitting TS

The input signal $x[t]$ is first split into P subsignals $x_p[t], p \in [0, P-1]$ by a transience splitter (TS). These signals correspond to P degrees of transience. Signal $x_{P-1}[t]$ corresponds to the most transient components, and $x_0[t]$ to the non-transient, or steady components. Obviously, $x[t] = \sum_{p=0}^{P-1} x_p[t]$ holds. Note that in Figure 1, $P = 3$.

A common technique to implement a TS is to use a cascade of transient extractors. Each transient extractor uses a lower threshold than the previous one, and sends the non-transient signal to the next one.

Transient detection and extraction is a complex task on its own and several approaches exist [12], [13], [14]. The transient detection of the presented algorithm is based on previous work by the main author [6].

4.2. Magnitude analysis MA

In a second step, P magnitude analyzers (MA) extract the magnitude spectrums $|X_p[s,k]|$ of the P signals $x_p[t]$. The MAs use short-time Fourier transforms with different resolutions, higher time resolution being used for more

transient signals. Possible values for the STFT resolution τ for $P = 3$ are $\tau = 96$ ms for MA_0, $\tau = 24$ ms for MA_1 and $\tau = 6$ ms for MA_2.

4.3. Master-slave phase vocoder MS-PV

The original signal $x[t]$ and the signals $x_p[t]$ are then processed using a master-slave phase vocoder (MS-PV), consisting of a master part and a slave part.

The master part uses a phase vocoder (PV) to transform $x[t]$ into $v[t]$. Although it involves complex calculations, a PV ultimately results in an output spectrum $V[s,k] = X[s,k]\Phi_{s,k}$, where the complex coefficients $\Phi_{s,k}$ are determined by the exact algorithm used [1].

The slave part simply reuses the $\Phi_{s,k}$ coefficients computed by the master part. It transforms each $x_p[t]$ signal in the frequency domain using point by point multiplication by the $\Phi_{s,k}$ coefficients, giving the $v_p[t]$ signals.

$P + 1$ STFTs are used, one for $x[t]$ and one for each $x_p[t]$. Unlike other parts of the algorithm, they all use the same high frequency resolution that matches the resolution of MA_0 (the resolution used for steady sounds).

In other words, the MS-PV uses $x[t]$ to *calculate* the time-stretching transformation, and then it *applies* the transformation to every signal $x_p[t]$.

This technique ensures that the phases of the resulting signals $v_p[t]$ are coherent and that $v[t] \approx \sum_{p=0}^{P-1} v_p[t]$. Note that $v[t]$ itself is not further used in the workflow.

The master-slave technique is not new, although not necessarily named as such. It is commonly used to preserve the phase coherence between the left and right channels when processing stereo signals [4], [6]. In such a case, the master uses a monophonic signal that is the sum (or a more elaborate aggregate) of the left and right signals, and the slaves process the left and right signals themselves.

4.4. Magnitude-only phase vocoder MoPV

In the meanwhile, the magnitude spectrums $|X_p[s,k]|$ are processed through magnitude-only phase vocoders (MoPV), resulting in time-stretched magnitude spectrums $|W_p[s,k]|$. Each MoPV uses the time-frequency resolution of the corresponding MA. The MoPVs are trivially implemented by using different analysis and synthesis hop sizes.

4.5. Magnitude correction MC

Finally, the signals $v_p[t]$ are processed by magnitude corrections (MC) stages, using the magnitude spectrums $|W_p[s,k]|$. MCs are implemented in the frequency domain using STFTs. Again, each MC uses the time-frequency resolution of the corresponding MoPV and MA.

The magnitude correction step MC_p computes the spectrum $Y_p[s,k]$ as follows:

$$Y_p[s,k] = |W_p[s,k]|\, e^{i\, arg(V_p[s,k])}$$

where $arg(V_p[s,k])$ is the phase of $V_p[s,k]$, $V_p[s,k]$ is the STFT of the $v_p[t]$ signal, $|W_p[s,k]|$ is the magnitude spectrum transformed by $MoPV_p$, and $Y_p[s,k]$ is the STFT of the resulting signal $y_p[t]$.

In other words, each MC basically keeps the *phase* of $v_p[t]$, but replaces its *magnitude* by $|W_p[s,k]|$.

Before this step, all $v_p[t]$ signals are smeared due to the high frequency resolution used by the MS-PV. This magnitude correction step is essential as it precisely removes the smearing, proportionally to the transience of the signal. As the phases of $v_p[t]$ are untouched by this step, the resulting signals $y_p[t]$ are still phase coherent.

In the last step, the signals $y_p[t]$ are added to produce the final signal $y[t]$.

Observe that the MA, MoPV and MC steps use multiple, adaptive resolutions whereas the MS-PV and its PV use a single, high frequency resolution.

5. CONSISTENCY

As $v_0[t]$ and $|W_0[s,k]|$ are obtained using the same high frequency resolution, it may seem that the MA_0, $MoPV_0$ and MC_0 steps are useless. Why do we need to fix the magnitudes of $v_0[t]$ using the magnitudes $|W_0[s,k]|$ processed with the same time-frequency resolution?

While it is possible to remove these steps, measurements revealed that it results in an attenuation of up to 3 dB of the amplitude of atonal signals such as white noise. Tonal signals on the other hands are not affected. Hence the presence of these steps helps in better keeping the balance between atonal and tonal components.

The reason of these attenuations is that horizontal phase coherence cannot be preserved for noise-like signals that have, by definition, chaotic phases. This is an instance of the more general "consistency" problem [1], [15]. This problem is significantly reduced by the proposed technique in a simple way, even for the degenerated case of $P = 1$.

6. FURTHER IMPROVEMENTS

6.1. Stereo handling

The input signal $x[t]$ was assumed to be monophonic. In practice it is advisable to also handle stereo signals consisting of a left channel $x^L[t]$ and a right channel $x^R[t]$.

This can be done by duplicating every signals, magnitude spectrums and processes in the workflow of Figure 1, with the exception of the master-slave phase vocoder (MS-PV). Instead of duplicating it into one instance that uses $x^L[t]$ for the master part and another instance that uses $x^R[t]$, a single instance is used. It uses $x_L[t] + x_R[t]$ for the master part (or any more elaborate aggregate), and applies the transformation to all the $x_p^L[t]$ and $x_p^R[t]$ signals.

This preserves phase coherence not only between components of different transience, but also between the left and right channels. The technique can easily be extended to signals with more channels, such as surround sounds.

6.2. Pyramidal resolutions

The presented technique uses different time-frequency resolutions, which are chosen adaptively based on an analysis of the signal transience. It is possible to combine it with the technique of pyramidal resolutions that just gives lower frequency resolution for high frequencies and higher frequency resolution for low frequencies [4], [8].

Preliminary tests showed audible improvements due to the fact that using pyramidal resolutions seems to be complementary with the adaptive technique:

- Strong transients are not processed well with pyramidal resolutions alone (they are still smeared, although a bit less), but are generally well detected and processed by the presented adaptive technique
- Transients that are too weak to be detected by the presented adaptive technique (such as phase-only transients) are generally well processed using pyramidal resolutions.

6.3. Other improvements

Finally, the following improvements are currently investigated:

- Resetting the phase at transients [3]. This is not incompatible with the presented approach and might improve it for extreme stretching factors.
- Inclusion of techniques for phasiness reduction [16], [17].
- Better handling of noise [18], [19].
- Investigation of other transience splitting schemes [12], [13], [14].

7. RESULTS

The presented technique has been implemented in Java, with a choice of $P = 3$. Figure 2 shows spectrograms comparing the presented technique with a phase vocoder. A time stretching factor of 1.5 was used. The spectrograms use 1024 point FFTs at 44100 Hz. The frequency range is 0 to 16 kHz, on a logarithmical scale. The signal duration is 1 s for (a) and 1.5 s for (b) and (c). Additional audio samples, comparisons and the Java implementation as a command-line tool are available on-line [20].

Figure 2 shows that the proposed technique does not smear the transients (visible as vertical lines), unlike the unmodified phase vocoder. This holds despite of the presence of steady tones (horizontal lines) and noise-like signals (large grey areas) at the same time.

Informal listening tests confirmed these observations. Transients are well preserved and non-transient components are not affected, even when both occur at the same time.

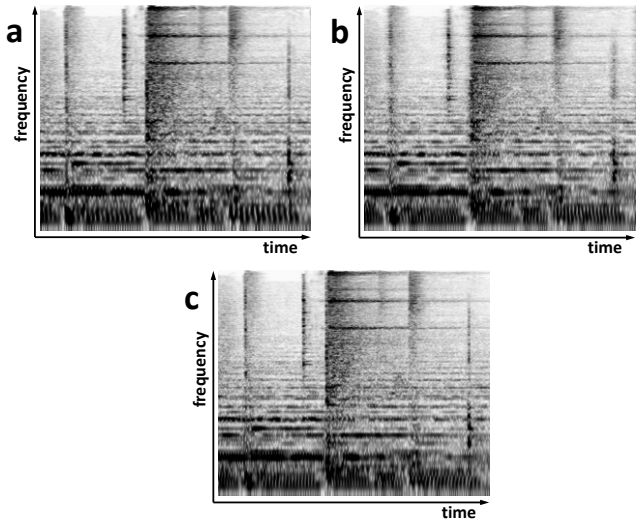

Figure 2: Spectrograms of: (a) polyphonic music signal, (b) time stretched with a phase vocoder, (c) time stretched with the proposed technique.

Other techniques that shift transients in time or reset the phase tend to either leave a residual smearing, or to introduce a mechanical, "hashed" timbre in similar situations. The reason is that they treat like a transient any non-transient components that occur at the same time.

However, when transients are stretched by large ratios ($\alpha < 0.5$ or $\alpha > 2$), they sometimes sound unnatural with the proposed technique, precisely because they are stretched like the rest of the signal. Techniques that just shift them in time might give better results in these extreme cases.

Note that on the other hand, when pitch shifting is done by combining time stretching with resampling, the presented technique always preserves the duration of the transients.

8. CONCLUSION

An adaptive time-frequency resolution technique for time stretching has been proposed. Based on an analysis of the transience of the signal in both time and frequency, it stretches transients with higher time resolution instead of just shifting them in time. No smearing artefacts are introduced for moderate time stretching factors, and non-transient components are better preserved. Phase coherence problems are coped with by processing the phase with a single resolution, and using multiresolution for the magnitude. The proposed technique also better preserves the balance between tonal and atonal components. Apart from the analysis of the transience of the signal, the actual processing only requires short-time Fourier transforms, a phase vocoder, and basic arithmetic.

9. REFERENCES

[1] Jean Laroche and Mark Dolson, "Improved Phase Vocoder Time-Scale Modification of Audio", *IEEE Transactions on Speech and Signal Processing*, vol. 7, no. 3, May 1999.

[2] Frederik Nagel and Andreas Walther, "A Novel Transient Handling Scheme for Time Stretching Algorithms", *127th Audio Engineering Society Convention*, New York, pp. 185-192, 2009.

[3] Axel Röbel, "A New Approach to Transient Processing in the Phase Vocoder", *proc. of the 6th intl. conf. on Digital Audio Effects*, London, UK, September 2003.

[4] Jordi Bonada, "Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio", *International Computer Music Conference*, pp. 396-399, 2000.

[5] Alexey Lukin and Jeremy Todd, "Adaptive Time-Frequency Resolution for Analysis and Processing of Audio", *120th Audio Engineering Society Convention*, Paris, France, May 2006.

[6] Nicolas Juillerat, Stefan Müller Arisona and Simon Schubiger-Banz, "Enhancing the Quality of Audio Transformations Using the Multi-Scale Short-Time Fourier Transform", *proc. of the 10th IASTED conf. on Signal and Image Processing*, Hawai, USA, August 2008.

[7] Jonathan Driedger and Meinard Müller, "A Review of Time-Scale Modification of Music Signals", *Applied Science*, vol. 6, issue 2, pp. 57, 2016.

[8] Nicolas Juillerat, Stefan Müller Arisona, Simon Schubiger-Banz, "A Hybrid Time and Frequency Domain Audio Pitch Shifting Algorithm", *proc. of the 125th Audio Engineering Society Convention*, San Francisco, CA, October 2008.

[9] Tony S. Verma and Teresa H.Y. Meng, "Time Scale Modification Using a Sines+Transients+Noise Signal Model", *proc. of the Digital Audio Effects Workshop*, Barcelona, pp. 49-52, 1998.

[10] Marco Liuni, Axel Röbel and Ewa Matusiak, "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, 2013.

[11] N. Chintaiah, "Time Scale Modification of Speech Signals using Wavelet Packet Transform", *Journal of Science and Research*, vol. 4, issue 12, pp. 1174-1179, 2015.

[12] Laurent Daudet, "A Review on Techniques for the Extraction of Transients in Musical Signals", *proc. of the 3rd intl. conf. on Computer Music Modeling and Retrieval*, Springer Verlag, pp. 219-232, 2005.

[13] Francisco Jesus Canadas-Quesada et al. "Percusssive / harmonic sound separation by non-negative matrix factorization with smoothness / sparseness constraints", *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2014, no. 26, July 2014.

[14] Derry FitzGerald, "Harmonic/Percussive Separation using Median Filtering", *proc. of the 13th intl. conf on Digital Audio Effects*, Graz, Austria, September 2010.

[15] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236-243, April 1984.

[16] David Dorran, Eugene Coyle and Robert Lawlor, "Audio Time-Scale Modification using a Hybrid Time-Frequency Domain Approach", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2005.

[17] Sebastian Kraft, Martin Holters, Adrian von dem Knesebeck and Udo Zölzer, "Improved PVSOLA Time-Stretching and Pitch-Shifting for Polyphonic Audio", *proc. of the 15th intl. conf. on Digital Audio Effects*, York, UK, September 2012.

[18] Wei-Hsiang Liao, Axel Roebel and Alvin W.Y. Su, "On Stretching Gaussian Noise with the Phase Vocoder", *proc. of the 15th intl. conf. on Digital Audio Effects*, York, UK, September 2012.

[19] Alexis Moinet, Thierry Dutoit and Philippe Latour, "Audio Time-Scaling for Slow Motion Sports Video", *proc. of the 16th intl. conf. on Digital Audio Effects*, Maynooth, Ireland, September 2013.

[20] Nicolas Juillerat, "Audio Time Stretching with an Adaptive Multiresolution Phase Vocoder: Companion website", available online at: http://www.pitchtech.ch/Confs/ICASSP2017/index.html