TIME-SCALING OF AUDIO SIGNALS WITH MUTI-SCALE GABOR ANALYSIS

Olivier Derrien

ISITV - Université du Sud Toulon-Var
Av. G. Pompidou, BP 56 F-83162, La Valette du Var Cedex, France
olivier.derrien@univ-tln.fr

ABSTRACT

The phase vocoder is a standard frequency domain time-scaling technique suitable for polyphonic audio, but it generates annoying artifacts called phasiness, or loss of presence, and transient smearing, especially for high values of the time-scale parameter. In this paper, a new time-scaling algorithm for polyphonic audio signals is described. It uses a multi-scale Gabor analysis for low-frequency content and a vocoder with phase-locking on transients for the residual signal and for high-frequency content. Compared to a phase-locking vocoder alone, our method significantly reduces both phasiness and transient smearing, especially for high values of the time-scale parameter. For time-contraction (time-scale parameters lower that one), the results seem to be more signal-dependant.

1. INTRODUCTION

Time-scale modification of audio aims at changing the playback rate of a recorded signal without altering its frequency content, i.e. pitch and timbre. For instance, time-scaling is useful for electronic music composers who want to synchronize musical samples in order to produce a coherent output signal. A time-scaling effect consists either of a speeding up, called time-contraction, either of a slowing-down, called time-stretching.

Time-scaling techniques can be roughly classified in two categories: time-domain and frequency-domain. Time domain algorithms, typically synchronized overlap-add (SOLA) [1], are usually very efficient and can produce high-quality audio output, but only when applied to quasi-periodic signals, speech for instance. In the case of more complex audio content, like polyphonic music, time-domain methods perform poorly. Frequency-domain methods, typically phase vocoder, can be applied to both quasi-periodic and complex audio signals, still with major drawbacks: a higher computational cost and annoying artifacts in the output signal. These artifacts are usually known as transient smearing and phasiness. Transient smearing consists of a loss of percussiveness, and phasiness can be compared to an artificial reverberation effect, or a loss of presence. In fact, these two aspects are related: smooth attacks and a notable reverberation are often associated with a long distance between the source and the listener.

The phase vocoder was introduced by Flanagan *et al.* [2] in 1966, but a considerable amount of studies have focused on improving the vocoder audio quality. Laroche *et al.* [3] explained the phasiness effect by a loss of phase consistency across the vocoder channels, and developed a phase locking technique to restore partially this coherence. This method can be considered as an improvement of the method by Puckette [4]. A constant frame-rate version of the phase vocoder was proposed by Bonada [5]. Different phase-locking techniques on transients location were published by Duxbury *et al.* [6], and by Röbel [7]. Dorran *et al.* [8] also proposed a method for maintaining phase coherence between vocoder channels, but only for moderate time-scale factors. A real-time software implementation was recently described by Karrer *et al.* [9] and an hybrid approach mixing SOLA and phase vocoder was proposed by Dorran *et al.* [10]. Despite significant improvements, some artifacts remain.

Sinusoidal modeling is another class of frequency techniques suitable for time-scaling of audio. More precisely, sinusoidal modeling is commonly used in parametric audio/speech coding at low bitrate, for instance in MPEG-4 HILN [11]. The output of the synthesis module can be easily time-scaled, but the overall signal quality is poor (typically between 1/5 and 2/5 on the MOS scale [12]). Surprisingly, sinusoidal modeling for high-quality time-scaling of audio signals have received very few attention so far. In this paper, we describe a new time-scaling technique based on a multi-scale sinusoidal analysis. We also propose a hybrid time-scaling algorithm combining this method to a phase-locking vocoder, and show that both transient smearing and phasiness are significantly reduced compared to the phase-locking vocoder alone.

The paper is organized as follows: section 2 provides an overview of the phase vocoder with phase-locking on transients. In section 3, the focus is on our multi-scale sinusoidal analysis and its application to time-scaling of audio signals. Section 4 describes the hybrid algorithm and a comparison with the vocoder alone is given. Section 5 concludes.

2. PHASE VOCODER TIME-SCALING

In this section, we describe the phase vocoder that we have implemented as a reference method. Although it might not be considered as a top-level vocoder, the phase-locking technique significantly improves the signal quality compared to a basic phase vocoder.

2.1. Phase vocoder basics

In a Discrete Fourier Transform (DFT) implementation of the phase vocoder, the audio signal x is analyzed with a N-point DFT and a R_a hop-size. Thus, two successive analysis intervals overlap by $N - R_a$ samples. X are the DFT coefficients:

$$X(u,k) = \sum_{n=0}^{N-1} w_a[n]\, x[n + uR_a - N/2]\, e^{-j2\pi\frac{kn}{N}} \qquad (1)$$

w_a is the analysis window, $u \in \mathbb{N}$ is the analysis interval index, and $k \in [0 \cdots N-1]$ is a frequency index. Each value of index k corresponds to a vocoder channel. uR_a are the analysis time-instants.

Between the analysis and the synthesis stage, the signal is modified in the DFT domain. These modifications will be explained further on. Y denote the modified coefficients. The synthesis involves an iDFT:

$$y_u[n] = \frac{1}{N} \sum_{k=0}^{N-1} Y(u,k)\, e^{j2\pi \frac{kn}{N}} \tag{2}$$

for $n \in [0 \cdots N-1]$, $y_u[n] = 0$ otherwise. The final output signal y is obtained with an overlap-add operation:

$$y[n] = \sum_u y_u[n - uR_s + N/2] \tag{3}$$

R_s is the synthesis hop-size, and uR_s are the synthesis time-instants. The time-scale factor is: $\alpha = \frac{R_s}{R_a}$.

In the absence of modification, i.e. $\alpha = 1$, one simply define $Y(u,k) = X(u,k)$, and the output signal y is similar to x, depending on the analysis window w_a. For instance, a Hanning window ensures the perfect reconstruction. When $\alpha \neq 1$, the amplitude of the DFT coefficients is preserved: $|Y(u,k)| = |X(u,k)|$, but the phases are modified according to the following method.

At the first analysis/synthesis instant, we initialize:

$$\angle Y(0,k) = \angle X(0,k) \tag{4}$$

Other initializations are possible, but this one suits any value of the time-scale factor α [3]. If the signal in each channel were a single pure sine of frequency $2\pi\frac{k}{N}$, the modified phase $\angle Y(u,k)$ could be computed for every u according to the phase propagation formula from instant $(u-1)R_s$ to uR_s:

$$\angle Y(u,k) = \angle Y(u-1,k) + R_s\, 2\pi\frac{k}{N} \tag{5}$$

However, the signal is not a single pure sine, and the DFT coefficients exhibit a phase increment. The analysis phase increment can be measured:

$$\Phi_a(u,k) = \angle X(u,k) - \left(\angle X(u-1,k) + R_a\, 2\pi\frac{k}{N}\right) \tag{6}$$

The synthesis phase increment is $\Phi_s(u,k) = \alpha\, \mathrm{PD}\left(\Phi_a(u,k)\right)$, where PD is the principal determination of an angle. Finally, the complete phase propagation formula is:

$$\angle Y(u,k) = \angle Y(u-1,k) + R_s\, 2\pi\frac{k}{N} + \Phi_s(u,k) \tag{7}$$

2.2. Phase locking at transient locations

Computing the synthesis phases according to the phase propagation formula (7) ensures the horizontal phase coherence inside each channel. But the vertical phase coherence between channels is lost, which causes transient smearing and phasiness [3].

Obviously, both horizontal and vertical phase coherence can not be achieved at any time and for every channel, but many researches have focused on finding a good balance between the two. Recent studies have shown that the vertical coherence is particularly crucial at transient locations [5, 6, 7]. Thus, preserving the horizontal phase coherence on stationary regions and forcing vertical coherence at transients, for instance by resetting the synthesis phases, also called phase-locking, seems to be a good solution, but it requires a transient detection algorithm. However, resetting the phases on high-energy stationary partials coming though a transient region must be avoided, because the signal energy suddenly collapses in front of the transient. In solution proposed by Duxbury *et al* [6], only the stationary regions are time-scaled, whilst the phase is locked and the time-scale factor is forced to be one at transients. Despite local variances in time-scaling factor, rhythm is maintained globally. In the algorithm by Röbel [7], both the transient detection and the transient processing algorithms operate on the level of frequency channels: the transient detection process classifies the channels in transient/non-transient content, and the synthesis phase is reset only in non-transient channels. Furthermore, the phase reset is performed only when the transient is located close to the center of the analysis interval, so there is no need to force the time-scale factor to be one.

2.3. Implementation details

The phase-locking vocoder that we implemented as a reference technique is close to algorithm proposed by Röbel.

The choice of the DFT size N is a trade-off between frequency-distortion on low-frequency partials and transient smearing: a high value for N gives a good ability to reproduce low-frequency partials but generates a considerable transient smearing effect. At $f_s = 44100$ Hz, 2048 samples (46.5 ms) seems to be a good value. The choice of the analysis hop-size R_a is a trade-off between high-frequency buzzy artifacts due to the synthesis overlap-add, and a coarse discretization step for the time-scaling factor α: a high value for R_a produces a high quality synthesis, but as $R_s \in \mathbb{N}^*$, the possible time-scaling factors are $\alpha = \frac{k}{R_a}, k \in \mathbb{N}^*$. If we set $R_a = 8$ samples at $f_s = 44100$ Hz, synthesis artifacts are clearly perceptible for $\alpha = 1.5$. $R_a = 4$ samples seems to be a good value. Possible time-scaling factors are then 0.25, 0.5, 0.75, 1, 1.25, 1.5 etc.

The transient detection algorithm is based on the energy evolution in frequency subband, whilst Röbel proposes a more complex criterion (center of gravity of the instantaneous energy in each subband and each analysis interval). The signal, sampled at $f_s = 44100$ Hz, is analyzed with a 512-point DFT and a 75% overlap, to preserve a good time-resolution. In each subband, when the energy increases by more than 10 dB, the subband is marked. In each analysis interval, if the number of marked subbands exceeds half of the total number of subbands, we decide that a transient is located at the center of the interval. In the vocoder, the synthesis phases are reset only on marked subbands at transient-marked locations. On figure 1, we plot the spectrogram of a glockenspiel signal (from the SQAM database [13]), and phase-reset locations. One can observe that the high energy partials are preserved.

3. MULTI-SCALE GABOR ANALYSIS

In this section, we present the multi-scale sinusoidal analysis that we use in our time-scaling algorithm. First, we describe the redundant time-frequency dictionary composed of Gabor waveforms and the decomposition method which is basically a modified version on the Matching Pursuit algorithm. Then, we explain how the time-scaling operation is applied to each atom.

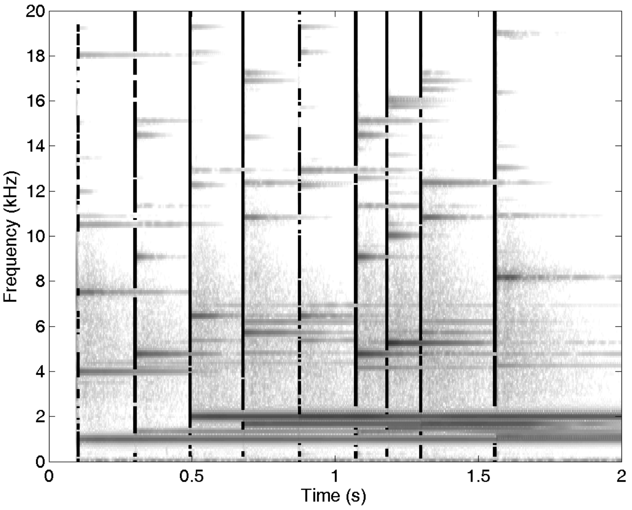

Figure 1: *Spectrogram of a glockenspiel signal (gray) and phase-reset locations (black).*

3.1. Time-frequency dictionary

The underlying signal model is a linear combination of time-frequency waveforms g plus a residual signal r:

$$x[n] = \sum_i a_i g_{\lambda_i}[n] + r[n] \tag{8}$$

$a_i g_{\lambda_i}[n]$ are called atoms. $g_\lambda[n]$ are complex Gabor waveforms [14], defined by:

$$g_\lambda[n] = \gamma(s)\, h_g\left(\frac{n-p}{s}\right) e^{j2\pi\nu n}, \quad \lambda = \{s, p, \nu\} \tag{9}$$

s is the time-scale factor, p the translation parameter and ν the modulation frequency. $h_g(t)$ is the amplitude function and $\gamma(s)$ is a normalization factor, depending on s. The dictionary is the overcomplete set of all possible waveforms. In a classic Gabor dictionary, $h_g(t)$ is a Gaussian function. For this application, we rather use a Hanning window, which is a compactly-supported function:

$$h_g(t) = (1 + \cos(2\pi t))\ .\ \mathbf{1}_{[0,1[}(t) \tag{10}$$

Parameters are discretized in the following way:

$$\begin{aligned} s &= 2^q, \quad i \in \{q_{\min} \cdots q_{\max}\} & (11)\\ p &= uR_g, \quad u \in \mathbb{N} & (12)\\ \nu &= \frac{k}{s}, \quad k \in \{1 \cdots s-1\} & (13) \end{aligned}$$

R_g, the hop-size, is set to $2^{q_{\min}-1}$ and does not depend on the time-scale. This differs from the usual discretization in Gabor dictionaries, where the hop-size depends on the time-scale factor (usually $p = u\frac{s}{2}$). In other words, the overlap factor increases with the time-scale in our dictionary and equals 50% for $s = 2^{q_{\min}}$, whilst the overlap factor equals 50% for all time-scales in the usual discretization. This choice was made in order to limit the phase rotation between consecutive atoms, which is crucial in the context of time-scaling.

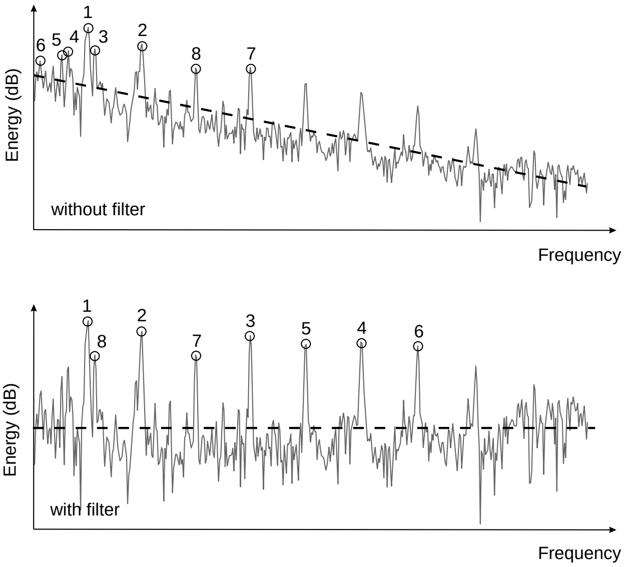

Figure 2: *Example of components selection order with Matching Pursuit in the frequency domain, with and without the adaptive filter.*

3.2. Decomposition algorithm

The decomposition algorithm determines a suitable set of index λ_i under a matching constraint, usually related to the energy of the residual signal r. The decomposition is performed on a frame-by-frame basis. Thus, only a limited subset of waveforms is considered in each frame. The time-segmentation stage is very similar to the one performed before a DFT: we use N-points intervals, with a R_a hop-size, and an analysis window w_a. We choose a set of parameters that match the Gabor dictionary: $N = 2^{q_{\max}}$, $R_a = R_g = 2^{q_{\min}-1}$ and $w_a[n] = h_g(\frac{n}{N})$. In the current frame, only the waveforms that completely overlap the analysis window are considered for the decomposition. When the same waveform is selected in different overlapping frames, which is a usual case, the final atom is computed by simply adding all the complex coefficients a_i associated to this waveform, bearing in mind the phase offset due to the translation of the analysis interval.

Our algorithm is a modified version of the iterative Matching Pursuit (MP) proposed by Mallat *et al.* [15]. The MP can be summarized as follows: at the beginning, the residual signal is equal to the signal itself. At each step, an atom is subtracted from the residual signal. This atom is co-linear to the waveform that maximizes the modulus of the inner-product with the residual signal. The decomposition is stopped when a matching criterion is smaller that a pre-defined threshold. The difference between the standard MP and our modified algorithm is that ours selects each atom from a filtered version of the residual signal. The filter transfer function is log-linear and computed for each frame so that the baseline of the filtered signal spectrum is approximately flat. Without this filter, the standard MP algorithm picks the most energetic component in the residual signal at each iteration. For instance, a high-energy noise component in low-frequency will be selected before a high-frequency partial with a lower energy. With the filter, the

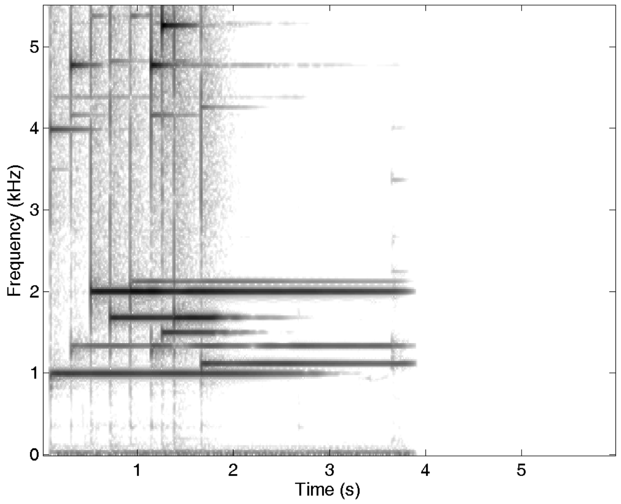

Figure 3: *Spectrogram of a glockenspiel signal.*

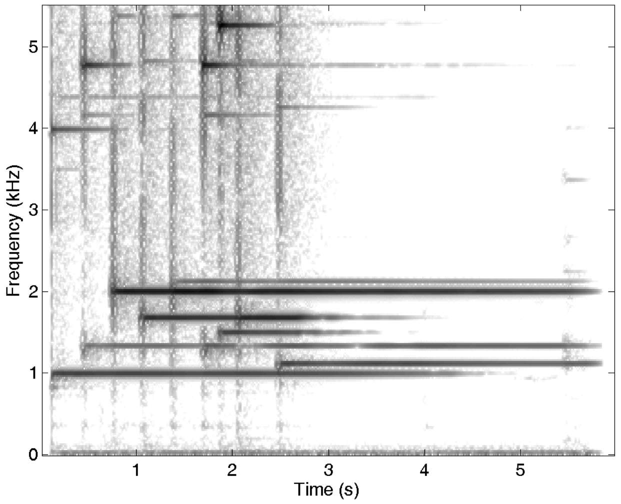

Figure 4: *Spectrogram of a glockenspiel signal time-scaled by a phase vocoder with phase-locking at transients, $\alpha = 1.5$.*

high-frequency partial is amplified and selected before the noise component. Our method improves the selection of significant partials, and leaves the noise components in the residual signal (see example on figure 2).

In the standard MP, the matching criterion is the energy of the residual signal. However, we found that combining this criterion with the correlation between the selected atom and the residual signal is more efficient. The exact description of our algorithm is as following. We denote $\tilde{x}$ and $\tilde{r}$ respectively the filtered versions of x and r, and M the matching criterion.

```
Initialization : set $i = 0$, $r_0 = x$ and $\tilde{r}_0 = \tilde{x}$
while $M(\tilde{r}_i) > \varepsilon$
    Compute $\forall\lambda$ the inner-product $\langle \tilde{r}_i, g_\lambda \rangle$
    Select the best waveform index:
        $\lambda_i = \text{Argmax}_\lambda |\langle \tilde{r}_i, g_\lambda \rangle|$
    Subtract the corresponding atom:
        $a_i = \langle r_i, g_{\lambda_i} \rangle$
        $\tilde{a}_i = \langle \tilde{r}_i, g_{\lambda_i} \rangle$
        $r_{i+1} = r_i - a_i \, g_{\lambda_i}$
        $\tilde{r}_{i+1} = \tilde{r}_i - \tilde{a}_i \, g_{\lambda_i}$
    Increment the waveform index: $i = i + 1$
end
```

3.3. Atoms time-scaling

Assuming that the residual signal is not perceptually significant, the time-scaling operation can be achieved by scaling the linear combination of time-frequency waveforms, i.e. by scaling each atom. The basic rule for scaling an atom is as follows: on stationary regions, the time-scale parameter s and the translation parameter p are scaled, whilst the modulation frequency ν remains unchanged. When the center of an atom is located on a transient, the atom is not scaled in order to preserve the time-envelope of the transient.

Concerning amplitude and phase of the modified atoms, we propose the following rule: for the current atom, if no previous overlapping atom with the same frequency exists in the decomposition, the amplitude and phase are kept unchanged. Otherwise, the amplitude is kept unchanged and the phase propagation formula is applied.

More precisely: first, in the decomposition formula (8), the atoms are classified according to:

1. increasing translation parameter p
2. decreasing energy $|a_i|^2$

Then, for each atom $a_i g_{(s_i,p_i,\nu_i)}$, the modified atom $a'_i g_{(s'_i,p'_i,\nu_i)}$ is computed as follows. Concerning the waveform parameters:

- if p_i is located on a transient and if ν_i is in a transient-marked subband, $s'_i = s_i$ and $p'_i = p_i$.
- else, $s'_i = \alpha s_i$ and $p'_i = \alpha p_i$.

Concerning the complex coefficient, the amplitude is preserved: $|a'_i| = |a_i|$, and for the phase:

- if a previous overlapping atom $a_j g_{(s_j,p_j,\nu_j)}$ with $\nu_j = \nu_i$ exists in the decomposition, the phase is set according to the phase-propagation formula. The phase increment between atoms j and i is:

$$\Phi_{ji} = \angle a_i - (\angle a_j + (p_i - p_j) 2\pi\nu_i) \tag{14}$$

  and the modified phase is:

$$\angle a'_i = \angle a'_j + \alpha(p_i - p_j) 2\pi\nu_i + \alpha \text{PD}\left(\Phi_{ji}\right) \tag{15}$$

- else $\angle a'_i = \angle a_i$.

With this method, no explicit phase-locking is necessary on transient locations.

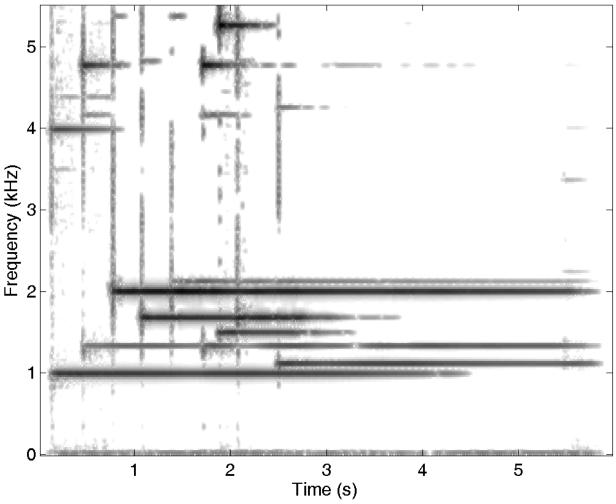

Figure 5: *Spectrogram of a glockenspiel signal time-scaled by Gabor analysis, $\alpha = 1.5$.*

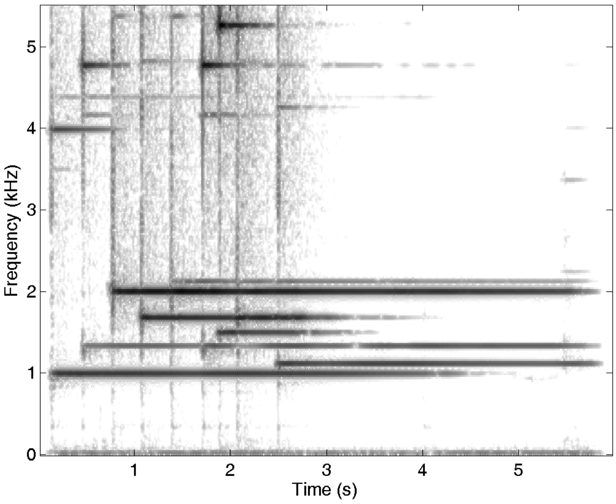

Figure 6: *Spectrogram of a glockenspiel signal time-scaled by the hybrid method, $\alpha = 1.5$.*

3.4. Implementation details and first results

According to Mallat [15], the complexity of the Matching Pursuit is similar to the one of the FFT i.e. $\mathcal{O}(N\log(N))$, when implemented in a efficient way. Practically, one can observe that the MP implementation is significantly more complex that the FFT. In our experiments, we chose to downsample the audio signal from 44100 Hz to 11025 Hz in order to limit the complexity.

The length of analysis intervals is set to $N = 1024$ samples (93 ms), which is twice the length of the phase vocoder analysis intervals. Thus, the theoretical frequency-resolution is twice better. The hop-size is set to $R_g = 64$ samples. We get $i_{\max} = 10$ and $i_{\min} = 7$. The theoretical time-resolution is 11.5 ms. However, the practical time and frequency resolution strongly depend on the decomposition algorithm.

We have tested both methods, phase vocoder and Gabor analysis, through informal listening test on real polyphonic music signals, for different time-scale factors. Concerning the phase vocoder, the main conclusions are:

- The phase locking technique significantly reduces artifacts,
- But perceptible phasiness and transient smearing effects still appear.

and for Gabor analysis:

- On downsampled signal, this method generates fewer artifacts than the phase-locking vocoder,
- But noise components are missing.

As a graphical illustration, we show spectrograms of a glockenspiel signal. Figure 3 corresponds to the original (unprocessed) signal. On figure 4, the signal is time-scaled with the phase-locking vocoder for $\alpha = 1.5$. The audible transient smearing effect is visually noticeable on this plot: attack regions are stretched and look granular. Otherwise, the frequency content of the original signal seems preserved. On figure 5, the signal is time-scaled with the Gabor analysis method, with a signal-to-residual noise around 30 dB. This corresponds to an average number of 35 atoms per frame of 1024 samples (about 20 atoms in stationary regions, and about 150 to 200 atoms in transient regions). One can see that the time-smearing effect is reduced, but only high energy components are treated, and most of the noise components are left in the residual signal.

4. HYBRID TIME-SCALING

In this section, we describe our complete time-scaling method, based on both Gabor analysis and phase vocoder.

4.1. Hybrid method

The Gabor analysis method can hardly by used alone for time-scaling a full-bandwidth audio signal, because it would require a very high number of atoms per frame, possibly higher than the number of samples, and the resulting complexity would be excessive. We think that the most efficient approach consists of stopping the Gabor analysis when no significant partial is left in the residual signal. It can be achieved by downsampling the original signal and perform the Gabor analysis with a medium matching criterion. The atoms are scaled according to the algorithm decribed in the previous section. The residual signal, which contains noise components in the low-frequency band and all the high-frequency content, is scaled with a phase-locking vocoder. As there is no significant partial left in the residual signal, one can choose a higher time-resolution than when scaling the full signal. We set $N = 1024$ (23 ms) and $R_a = 4$ samples. The transient smearing effect is not contained, and no buzzy artifact is perceptible.

4.2. Final results

On figure 6, we plot the spectrogram of the glockenspiel signal scaled with our hybrid Gabor analysis/vocoder technique. One can observe that, compared to the Gabor analysis alone, the transient smearing effect is not increased and remains lower than with vocoder alone, whilst noise components are preserved in the scaled signal.

Informal listening tests, involving 4 listeners and 4 different audio excerpts have shown that the signal quality is improved compared to the phase-locking vocoder alone. Our method significantly reduces both the transient smearing and phasiness effects: the presence effect in the scaled signal is much better than with the phase vocoder alone, especially for high values of the scaling parameter ($\alpha > 1.5$), for which the scaled signal often sounds artificial. However, when $\alpha < 1$, the phase vocoder might perform better, on some very specific audio signals.

Examples of audio signals processed with both methods for various scaling parameters can be found on the DESAM project website: http://www.tsi.enst.fr/~rbadeau/desam/spip.php?article16.

5. CONCLUSION

In this paper, a new high-quality time-scaling algorithm for polyphonic audio signals has been presented. It is based on a multiscale Gabor analysis for low-frequency content (between 0 and 5.5 kHz), and on a phase-locking vocoder for high-frequency content (between 5.5 and 22 kHz) and for the residual part in the low-frequency band. In the time stretching context, i.e. $\alpha > 1$, our method significantly reduces the two main artifacts generated by a phase vocoder: phasiness and transient smearing. The improvement is particularly interesting when the scaling parameter is high ($\alpha > 2$). However, for time-contraction, i.e. $\alpha < 1$, the results seem to be more signal-dependant.

Compared to the phase vocoder, the overall complexity of the time-scaling process is significantly higher with our method. First because the Gabor analysis is more complex than a FFT, second because our method also requires a phase vocoder for the residual signal. This makes out method unsuitable for real-time implementations for the moment.

This study proves that Gabor analysis is a valid alternative to the phase vocoder for audio time-stretching, but must be considered as preliminary. In further studies, we will extend our method to full-bandwidth signals. We will also try to define more complex rules for time-stretching the atoms, with partials tracking for instance. We also work on a more complex signal model which would not require a phase vocoder for processing the residual signal.

6. ACKNOWLEDGEMENTS

The research leading to this paper was supported by the French GIP ANR under contract ANR-06-JCJC-0027-01, Décomposition en Éléments Sonores et Applications Musicales - DESAM. See the project website: http://www.tsi.enst.fr/~rbadeau/desam/ for more details.

7. REFERENCES

[1] S. Roucos and A. M. Wilgus, "High quality time-scale modification of speech," in *Proc. of the IEEE ICASSP International Conference on Acoustics, Speech and Signal Processing*, 1985.

[2] J. L. Flanagan and R. M. Golden, "Phase vocoder," *The Bell System Technical Journal*, vol. 45, pp. 1493–1509, November 1966.

[3] 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, May 1999.

[4] M. S. Puckette, "Phase-locked vocoder," in *Proc. of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New-York*, 1995.

[5] J. Bonada, "Automatic technique in frequency domain for near-lossless time-scale modification of audio," in *Proc. of the ICMC International Computer Music Conference, Berlin, Germany*, 2000.

[6] C. Duxbury, M. Davies, and M. Sandler, "Improved time-scaling of musical audio using phase locking at transients," in *Proc. of the 112th Convention of the Audio Engineering Society, Munich, Germany*, 2002.

[7] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. of the 6th International Conference on Digital Audio Effects (DAFx-03), London, UK*, 2003.

[8] D. Dorran, E. Coyle, and R. Lawlor, "An efficient phasiness reduction technique for moderate audio time-scale modification," in *Proc. of the 7th International Conference on Digital Audio Effects (DAFx-04), Naples, Italy*, 2004.

[9] T. Karrer, E. Lee, and J. Borchers, "Phavorit: A phase vocoder for real-time interactive time-stretching," in *Proc. of the ICMC International Computer Music Conference, New-Orleans, USA*, 2006.

[10] D. Dorran, E. Coyle, and R. Lawlor, "Audio time-scale modification using a hybrid time-frequency domain approach," in *Proc. of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New-York*, 2005.

[11] International Organization for Standardization, *ISO/IEC 14496-3 (Information technology - Very low bitrate audio-visual coding - Part3: Audio)*, 1998.

[12] R. Heusdens, J. Jensen, W. Bastiaan Kleijn, V. Kot, O. A. Niamut, S. Van De Paar, N. H. Van Schijndel, and R. Vafin, "Bit-rate scalable intraframe sinusoïdal audio coding based on rate-distortion optimization," *Journal of the Audio Engineering Society*, vol. 54, no. 3, pp. 167–188, March 2006.

[13] EBU SQAM, "Sound quality assessment material recordings for subjective tests," Compact Disc, 1998.

[14] N. Delprat, B. Escudie, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and P. Torresani, "Asymptotic wavelets and gabor analysis: Extraction of instantaneous frequencies," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 644–664, March 1992.

[15] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.