ADAPTIVE SINUSOIDAL MODELING OF PERCUSSIVE MUSICAL INSTRUMENT SOUNDS

Marcelo Caetano[1], George P. Kafentzis[2,3], Athanasios Mouchtaris[1,2], Yannis Stylianou[1,2]

[1]Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Greece
[2]Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece
[3]Orange Labs, TECH/ACTS/MAS, Lannion, France

caetano@ics.forth.gr, kafentz@csd.uoc.gr, mouchtar@ics.forth.gr, styliano@ics.forth.gr

ABSTRACT

Percussive musical instrument sounds figure among the most challenging to model using sinusoids particularly due to the characteristic attack that features a sharp onset and transients. Attack transients present a highly nonstationary inharmonic behaviour that is very difficult to model with traditional sinusoidal models which use slowly varying sinusoids, commonly introducing an artifact known as *pre-echo*. In this work we use an adaptive sinusoidal model dubbed eaQHM to model percussive sounds from musical instruments such as plucked strings or percussion and investigate how eaQHM handles the sharp onsets and the nonstationary inharmonic nature of the attack transients. We show that adaptation renders a virtually perceptually identical sinusoidal representation of percussive sounds from different musical instruments, improving the Signal to Reconstruction Error Ratio (SRER) obtained with a traditional sinusoidal model. The result of a listening test revealed that the percussive sounds modeled with eaQHM were considered perceptually closer to the original sounds than their traditional-sinusoidal-modeled counterparts. Most listeners reported that they used the attack as cue.

Index Terms— musical instrument modeling, percussion musical instruments, adaptive sinusoidal models, attack transients, pre-echo

1. INTRODUCTION

Sinusoids have been used to model the quasi-periodic (stationary) oscillations in music and speech [1, 2, 3, 4]. However, the sounds of musical instruments and speech also contain transients and noise. Percussive sounds produced by plucking strings (such as harpsichords, harps, and the *pizzicato* playing technique) or striking percussion instruments (such as drums, idiophones, or the piano) are notoriously difficult to model because they feature a sharp attack with highly nonstationary oscillations that die out very quickly, called transients. A transient signal is essentially localized in time and it has a finite duration. Therefore, transients are poorly represented as stationary oscillations (i.e., by slowly-varying sinusoids). The attack is the most salient perceptual feature of musical instrument sounds that listeners use in dissimilarity judgments [5, 6, 7]. It is well known that much of the characteristic quality of many musical sounds derives from the characteristics of the attack [8], although the harmonic structure of the stationary oscillations (when there is a steady state) is also clearly important [9]. Thus modeling attack transients is essential to obtain a perceptually similar representation of musical instrument sounds. Modeling percussive musical instrument sounds with quasi-stationary sinusoids commonly results in perceptually less sharp attacks because of poor temporal resolution and an artifact know as pre-echo. Transients require shorter windows to be properly detected and modeled, but short windows blur stationary sinusoids in the frequency domain. Additionally, stationary sinusoids are supposed to vary slowly *inside* the analysis window, thus windows centered around the onset smear out the attack.

There have been different proposals to detect, model, and even use transients in instrument recognition, sound representation and transformation [10, 11, 12, 13]. Serra [3, 14] developed the deterministic plus stochastic model to account for the stationary and nonstationary characteristics of sounds. However, transients and noise are not modeled separately. Levine [13] proposes to decompose the sounds into sinusoids plus transients plus noise and model each separately. Short analysis windows increase temporal resolution to detect and model transients, therefore the use of multi-resolution techniques [15] seems like a natural choice to detect modulations at different time scales. Daudet [16] discusses the use of pruned wavelet trees in transient modeling due to the natural multiresolution nature of wavelets. Keiler *et al.* [17] propose to analyze transient musical instrument sounds with an auto-regressive model. Tan and Sen [18] propose to use the attack transient envelope in musical instrument recognition. Macon [19] uses an all-pole filter excited by an impulse to represent percussion musical instrument sounds such as woodblocks, xylophones, or chimes. Laroche [20] applies a multichannel excitation-filter model to piano sounds. Atomic decomposition algorithms [21, 22, 23] can render a sparse representation that is perceptually good provided that there are appropriate atoms in the dictionary. While the sparsity and potential quality of matching pursuit models are desirable, the atomic decomposition is a less natural representation of the physical process than sinusoidal oscillations. Naturally, each model has advantages and disadvantages. This work requires a musical instrument sound model that represents the physical process intuitively and compactly while rendering perceptually close representations.

In this work, we propose to represent the oscillatory modes of musical instruments with nonstationary sinudoids to capture both the quasi-stationary behavior and transients (as amplitude and frequency modulations). We use an adaptive sinusoidal model dubbed extended adaptive Quasi-Harmonic Model (eaQHM) [24] to represent percussive musical instrument sounds such as plucked strings and idiophones (struck). The eaQHM algorithm has been applied on the speech counterparts of percussive sounds, or stop sounds, outperforming standard sinusoidal models (SM) [25] with the same complexity (number of resynthesis parameters). The analysis stage uses one extra parameter (degree of freedom) to iteratively estimate parameter values. However, it is adaptation of the sinusoids *inside* the analysis window that allows representation of both transients and stationary oscillations with sharp onsets (no pre-echo) and very lit-

tle residual. We compare eaQHM and SM, evaluating the quality of the representations quantitatively and qualitatively, showing that the "signal to reconstruction error rate" (SRER) is higher for eaQHM locally (before the attack) and globally (whole duration). Then we confirm that eaQHM produces percussive sounds perceptually closer to the original recordings with a listening test.

The next section reviews eaQHM to show that complexity (fidelity of representation) lies in adaptation rather than the higher number of analysis degrees of freedom. Then, we compare the model representation of percussive musical instrument sounds with SM and eaQHM in terms of analysis and re-synthesis parameters (degrees of freedom). Next, we evaluate objectively and subjectively the SM and eaQHM model representation of percussive musical instrument sounds. Finally, we present the conclusions and discuss future perspectives of the work described.

2. ADAPTIVE SINUSOIDAL MODELING

In what follows, $x\left(t\right)$ represents the analysis step, $\hat{x}\left(t\right)$ represents the synthesis equation (the sinusoidal component), and $\bar{x}\left(t\right)$ is the subtractive residual $\bar{x}\left(t\right) = x\left(t\right) - \hat{x}\left(t\right)$. The Quasi-Harmonic Model (QHM) [26] liest at the heart of eaQHM, addressing frequency mismatch by frequency updating. QHM is defined as

$$x(t) = \left[\sum_{k=-K}^{K} \left(a_k + tb_k\right) e^{j2\pi \hat{f}_k t}\right] w\left(t\right) \tag{1}$$

where a_k and b_k are the complex amplitude and the complex slope of the k^{th} sinusoid and $w\left(t\right)$ is the analysis window. The term tb_k can be interpreted as the time-domain representation of the derivative in the frequency domain. The analysis frequencies $\hat{f}_k$ are initialized as harmonically related. There is a frequency mismatch between the true frequency f_k and the frequency $\hat{f}_k$ of the k^{th} sinusoid given by

$$\eta_k = f_k - \hat{f}_k. \tag{2}$$

The error η_k leads to the underestimation of amplitudes for sinusoidal models that rely on peak picking. Pantazis *et al.* [26] showed that QHM is able to provide an estimate of η_k given by

$$\hat{\eta}_k = \frac{1}{2\pi} \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2}. \tag{3}$$

QHM estimates the analysis parameters a_k and b_k by least-squared errors and uses them to obtain the synthesis parameters $|a_k|$ and $\hat{\eta}_k$. Thus the synthesized signal is represented as

$$\hat{x}(t) = \left[\sum_{k=-K}^{K} |a_k| e^{j\left(2\pi\left(\hat{f}_k + \hat{\eta}_k\right)t + \hat{\phi}\right)}\right] w\left(t\right). \tag{4}$$

Later, Pantazis *et al.* [27] suggested an adaptive QHM (aQHM) model that iteratively corrects the amplitude $|a_k|$ and frequency $(\hat{f}_k + \hat{\eta}_k)$ estimates. The aQHM algorithm improves the accuracy of parameter estimation, but it still uses stationary sinusoids *inside* the analysis window to represent the sounds. Therefore, the non-stationary oscillations such as transients cannot be well represented by QHM. Ideally, both the amplitudes and phases should capture variations that occur in time scales smaller than the size of the window like

$$\hat{x}(t) = \sum_{k=-K}^{K} \hat{a}_k\left(t\right) e^{j\hat{\phi}_k(t)} \tag{5}$$

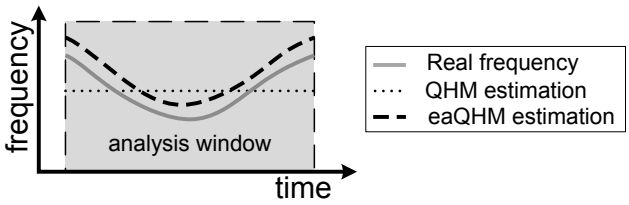

Fig. 1. Inside the analysis window, the frequency trajectory of a partial (solid grey line) is assumed to be constant for stationary sinusoidal models like QHM (dotted line), while eaQHM (dashed line) iteratively adapts to the shape of the instantaneous frequency.

where $\hat{a}_k\left(t\right)$ and $\hat{\phi}_k\left(t\right)$ are the instantaneous amplitude and phase of the k^{th} sinusoid. Kafentzis *et al.* [25] proposed to adapt both the instantaneous amplitude $\hat{a}_k\left(t\right)$ and the instantaneous phase $\hat{\phi}_k\left(t\right)$ inside the analysis window with the extended adaptive QHM (eaQHM). The analysis equation becomes

$$x(t) = \left[\sum_{k=-K_l}^{K_l} \left(a_k + tb_k\right) \hat{a}_k\left(t\right) e^{j\left(\hat{\phi}_k(t+t_l) - \hat{\phi}_k(t_l)\right)}\right] w(t) \tag{6}$$

where l is the frame index, a_k and b_k are the analysis parameters, and $\hat{a}_k\left(t\right)$ and $\hat{\phi}_k\left(t\right)$ are the time-varying amplitude and phase of each sinusoid respectively. The instantaneous values $\hat{a}_k\left(t\right)$ are estimated by spline interpolation inside the analysis window $w\left(t\right)$ and $\hat{\phi}_k\left(t\right)$ is the integral of the instantaneous frequency $\hat{f}_k\left(t\right)$, obtained by least squares interpolation. In this model, both $\hat{a}_k\left(t\right)$ and $\hat{\phi}_k\left(t\right)$ are iteratively adapted, so the signal is projected onto a set of non-stationary basis functions $\hat{a}_k\left(t\right) e^{j\hat{\phi}_k(t)}$ inside the analysis window that is locally adapted to the signal. Figure 1 illustrates the time-varying nature of the frequency trajectories for one sinusoid. The instantaneous amplitude can be depicted similarly.

Kafentzis *et al.* [24] showed that the basis functions are adapted to the local amplitude and phase characteristics of the signal, resulting in an adaptive AM-FM model of the analyzed signal. It was also shown that eaQHM can fully address the highly non-stationary nature of signals such as speech, both in its amplitude and in its phase. The adaptation algorithm can be found in [27]. The convergence criterion for eaQHM was the following:

$$\frac{\text{SRER}^{i-1} - \text{SRER}^{i}}{\text{SRER}^{i-1}} < \epsilon \tag{7}$$

where i is the iteration and SRER is the Signal-to-Reconstruction-Error Ratio of the resynthesized signal, defined as

$$\text{SRER} = 20\log_{10}\frac{\sigma_x}{\sigma_{x-\hat{x}}} = 20\log_{10}\frac{\text{RMS}\left(x\right)}{\text{RMS}\left(\bar{x}\right)} \tag{8}$$

where x is the original signal, $\hat{x}(t)$ is the reconstructed signal, and $\bar{x}$ is the residual signal, and σ_x denotes the standard deviation of x. Notice that $\sigma_x = \text{RMS}\left(x\right)$ because the waveforms have zero mean. In our experiments, ϵ is set to 0.01.

3. PERCUSSIVE MUSICAL INSTRUMENT SOUNDS AND ADAPTIVE SINUSOIDAL MODELING

The aim of this section is to illustrate the comparison between SM and eaQHM to foster the results of the evaluation. We want to show

that eaQHM represents percussive musical instrument sounds perceptually closer than SM with the same model complexity. Therefore, in this section we will show that adaptation gives sharp onsets without pre-echo with the same number of partials as SM. Figure 2 illustrates the ability eaQHM has to model sudden variations that happen in a time frame shorter than the hop size (inside the analysis window). In Figure 2 we see the reconstruction (resynthesized sound) zoomed in before the onset to show that the eaQHM representation is pre-echo free, while SM clearly presents smearing of the attack. The top of figure 2 shows the onset of the recording of a plucked guitar string. In the middle we see the waveform resynthesized from SM, and the bottom shows resynthesis with eaQHM. The pre-echo is highlighted in the middle, where the attack oscillations are also visually different from the original at the top. The next section measures these differences objectively (local SRER) and subjectively (listening test).

Adaptation also allows representation of transients as modulations of the quasi-stationary modes (sinusoidal partials) *inside* the analysis window. This results in each sinusoid representing more information than SM. Traditionally, SM parameter estimations are constant inside the analysis window, varying only between consecutive windows. The result is a sinusoidal representation that captures mostly quasi-stationary oscillations (that do not vary much inside the analysis window), leaving a residual that contains noise and transients. On the other hand, eaQHM leaves very little residual both before the onset (locally) and across the whole duration (globally). The evaluation will measure how much information is captured by each model (SM and eaQHM) locally and globally by comparing with the original signal.

The advantage that adaptation gives comes with the price of higher complexity in the analysis both for the number of parameters to be fit by the model and the iterative procedure. As evidenced by Section 2, eaQHM requires a more complex iterative analysis step, especially when compared with SM. However, once the eaQHM algorithm has fit the parameters of the model, the resynthesis model has the same complexity (eq. 5) as SM with a representation that is perceptually closer to the original recordings. Table 1 presents an overview of the analysis and synthesis complexity of SM and eaQHM to allow comparison. Complexity is considered as the number of parameters per frame each model requires to estimate (analysis) and represent (synthesis) K sinusoidal tracks. Notice that SM has the same complexity in the analysis and synthesis stages, while eaQHM fits more parameters iteratively (a few times until convergence) during the analysis stage than for resynthesis.

The next section evaluates the quality of the representation for both models with the same resynthesis complexity. Notice that eaQHM requires initialization of the frequency estimation procedure described in (2) and (3). For all the percussive sounds modeled with eaQHM in this work, we initialized the frequency estimation as 200 integer multiples of $f_0 = 40$ Hz. This limits the number

Table 1. Comparison of model complexity between SM and eaQHM for the analysis and synthesis stages. The table presents the number of parameters per frame as a function of the number of sinusoids K to estimate (analysis complexity) and to represent (synthesis complexity) sounds.

	Parameters per frame	
	SM	**eaQHM**
Analysis	$2K+1: a_k, f_k$	$3K+1: a_k, b_k, f_k$
Synthesis	$2K+1: a_k, f_k$	$2K+1: \hat{a}_k, \hat{\phi}_k$

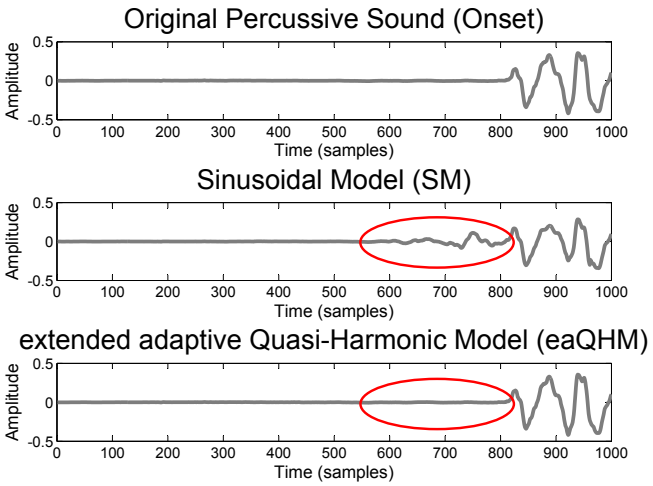

Fig. 2. Resynthesis around the onset for a percussive sound. The top shows the original waveform, the middle part illustrates SM while the bottom shows eaQHM.

of spectral peaks to 200 and spans the whole frequency range with the sampling frequency of $F_s = 16$ kHz used. For both SM and eaQHM, the Hann analysis window is set to 3 times the period of the lowest frequency f_0 (i.e., $T_0 = 20$ ms), a 2048-point FFT is computed for each analysis frame, and the hop size is set to 2 ms. For eaQHM, the maximum number of iterations is 5. Note that the algorithm might converge before reaching the maximum number of iterations if the convergence criterion in eq. (7) is met.

4. EVALUATION

In this section we present the results of the objective and subjective evaluation of the eaQHM model compared to SM. Both eaQHM and SM used the same input parameter values (number of sinusoids, window size, hop size, etc. See Section 3), differing only in the analysis complexity (see Table 1). The objective evaluation uses the Signal to Reconstruction Error Ratio (SRER). Higher values of SRER indicate less residual energy, therefore a better model representation. The subjective evaluation used was an online listening test to assess the perceptual similarity between the original recording of the percussive musical instrument sounds and the sounds resynthesized from the model parameters. The listening test aims to show that eaQHM renders sharper attacks (no pre-echo) and better modeled transients (nonstationarity), thus closer to the original recording with the same model complexity (number of resynthesis parameters) as SM.

Table 2 lists the percussive sounds used in this work [1] divided in two classes, *percussion* and *plucked strings*. All sounds belong to the same pitch class (C), ranging in pitch height from C3 ($f0 \simeq 131$ Hz) to C7 ($f0 \simeq 2093$ Hz), but most are C3 or C4. The dynamics range between *mezzo-forte* and *forte*, while the duration was kept under 2 s. The recordings were chosen to represent the range of percussive sounds with sharp attack and highly nonstationary transients from musical instruments commonly found in traditional Western orchestras. Two different *glockenspiel*, three different *vibraphone*, and three different *harp* sounds were used.

[1] Sounds from Vienna Symphonic Library database of musical instrument samples `http://www.vsl.co.at/en/65/71/84/1349.vsl`

4.1. Signal to Reconstruction Error Ratio

Table 3 shows both the *local* and *global* SRER values for eaQHM and SM calculated using (8). The *local* SRER is computed using only the samples in the first analysis window, right before the onset of the waveform. Therefore, the *local* SRER gives information about the presence or absence of *pre-echo*. Global values of SRER are computed for the whole waveform, serving as an estimate of the total residual energy "missed" by each model. Notice that eaQHM results in a higher SRER value both locally and globally.

4.2. Preference Listening Test

The aim of the listening test was to evaluate which resynthesised sound was perceptually closer to the original recording. We presented the original recording followed by resynthesis from SM or eaHQM (presented as model 1 and 2 in random order). The listener was asked to choose which was perceptually closer to the original recording. The test was forced choice, which means the listener did not have the option of selecting *no preference*. The test contained 19 percussive sounds, 13 plucked strings and 6 percussion instruments among those listed in Table 2. The listening test is available online at `http://www2.csd.uoc.gr/~kafentz/listest/pmwiki.php?n=Main.ListeningTest`. Table 4 shows the result of the listening test as percentage choice, that is, the percentage times each model was selected as perceptually closer to the original recording. In total, the results of 51 people were included in the evaluation.

4.3. Discussion

The result of the objective evaluation confirms that eaQHM presents a much higher value of both *global* and *local* SRER. Higher *global* SRER means that eaQHM represents more information from the original sound throughout, during nonstationary attack transients and more stationary oscillations. A higher *local* SRER indicates less energy around the onset, which means that eaQHM has virtually no pre-echo.

The result of the subjective evaluation, on the other hand, confirms conclusively that eaQHM renders percussive sounds perceptually closer to the original than a traditional SM. Table 4 shows that eaQHM outperformed SM in perceptual similarity for all sounds by at least 82%. All percussive sounds modeled with eaQHM were assessed perceptually similar to the original.

The participants could leave comments after taking the listening test. Most participants reported using the attack to tell the difference. A typical comment was *The beginning of the sound was the trigger for me. After the attack, it was more difficult to tell the difference.* Interestingly, some comments revealed clearly the strategy used to asses the perceptual similarity. For example, *To me, the sound quality of the pitched, sustained part was more or less identical. The onset, however, seemed to disturb the results for the simpler model in a short window around the onset.* Another listener reported that *The main difference I perceived was in the sharpness of the attacks.*

Table 2. Percussive sounds used in the listening test.

Percussion	Plucked String
marimba, glockenspiel, piano, xylophone, vibraphone, celesta	acoustic guitar, cello, classic guitar, harpsichord, harp, mandolin, ukelele, viola, violin

Table 3. Global and Local Signal to Reconstruction Error Ratio (SRER) values in dB for both models (SM and eaQHM) divided into two classes of percussive sounds, namely, percussion and plucked string.

	Global SRER (dB)			
	SM		eaQHM	
Percussion	$\mu = 16.65$	$\sigma = 2.55$	$\mu = 48.11$	$\sigma = 4.57$
Plucked String	$\mu = 19.46$	$\sigma = 4.92$	$\mu = 48.16$	$\sigma = 4.25$

	Local SRER (dB)			
	SM		eaQHM	
Percussion	$\mu = 12.31$	$\sigma = 2.82$	$\mu = 46.03$	$\sigma = 4.12$
Plucked Strings	$\mu = 13.40$	$\sigma = 4.14$	$\mu = 47.03$	$\sigma = 3.79$

For some of the shorter sounds the pitch also felt (very) slightly different, which seems to confirm that eaQHM does provide a different frequency estimation.

5. CONCLUSIONS AND FUTURE WORK

Percussive musical instrument sounds figure among the most challenging to model using sinusoids particularly due to the characteristic sharp onset and highly nonstationary nature of the attack transients. Traditional sinusoidal models fail to represent transients well with slowly-varying sinusoids and render a modeled sound whose onset is smeared in time (perceptually less sharp than the original) due to an artifact know as *pre-echo*. This paper proposes to model percussive sounds with an *adaptive sinusoidal model* due to its ability to accurately model sharp onsets and highly nonstationary attack transients. The extended adaptive QHM (eaQHM), which is a family member of adaptive sinusoidal models, is tested to confront this effect and it is shown that highly accurate, pre-echo-free representations of percussive sounds are possible using the adaptive approach. Results on a database of percussive sounds such as plucked strings and percussion instruments show that, on average, eaQHM improves by over 30 dB the Signal to Reconstruction Error Ratio (SRER) obtained by the standard sinusoidal model. A listening test showed that the percussive sounds modeled by eaQHM are perceptually closer to the original recordings than the same sounds represented by a traditional sinusoidal model for more than 80% of the listeners in all cases. Future perpectives include using eaQHM in transient detection and transient modeling for musical instrument recognition,

Table 4. Result of the listening test to assess the perceptual similarity between the original recording and the reconstructions with both models. The table shows in percentage how often eaQHM was selected perceptually closer to the original for each class of percussive sounds.

Percussion		**Plucked Strings**	
glockenspiel	98%	*acoustic guitar*	90%
glockenspiel 2	100%	*celesta*	92%
marimba	98%	*classic guitar*	98%
piano	86%	*harpsichord*	94%
vibraphone	98%	*harp*	96%
vibraphone 2	82%	*harp 2*	96%
vibraphone 3	98%	*harp 3*	92%
xylophone	98%	*mandolin*	92%
		ukelele	90%
		cello	90%
		violin	94%

segmentation, and sound transformations such as timbral variations, perceptually coherent time stretching and pitch shifting.

6. ACKNOWLEDGEMENTS

The authors would like to thank Gilles Degottex for his invaluable suggestions. This work is cofunded by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework-(NSRF), Research Funding Program: THALES, Project "MUSINET".

7. REFERENCES

[1] Yannis Stylianou, "Voice transformation," in *Springer Handbook of Speech Processing*, Jacob Benesty, M. Mohan Sondhi, and Yiteng Arden Huang, Eds., pp. 489–504. Springer, 2008.

[2] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, pp. 744–754, 1986.

[3] Xavier Serra and Julius O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 49–56, 1990.

[4] Xue Wen and Mark Sandler, "Source-filter modeling in the sinusoidal domain," *Journal of the Audio Engineering Society*, vol. 58, no. 10, 2010.

[5] John M. Grey and John W. Gordon, "Multidimensional perceptual scaling of musical timbre," *Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.

[6] E. L. Saldanha and John F. Corso, "Timbre cues and the identification of musical instruments," *The Journal of the Acoustical Society of America*, vol. 36, no. 11, pp. 2021–2026, 1964.

[7] Carol L. Krumhansl, "Why is musical timbre so hard to understand?," in *Structure and perception of electroacoustic sound and music*, S. Nielzén. and O. Olsson, Eds., pp. 43–54. Excerpta Medica, New York, 1989.

[8] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbre," *The Journal of the Acoustical Society of America*, vol. 94, pp. 2595, 1993.

[9] S. Handel, "Timbre perception and auditory object identification," in *Hearing*, B.C.J. Moore, Ed., pp. 425–461. Academic Press, New York, 1995.

[10] Marcelo Caetano, Juán Jusé Burred, and Xavier Rodet, "Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues," in *Proceedings of the International Conference on Digital Audio Effects*, 2010.

[11] Laurent Daudet, "A review on techniques for the extraction of transients in musical signals," in *Proceedings of the International Conference on Computer Music Modeling and Retrieval*, 2006.

[12] R. Bader, "Theoretical framework for initial transient and steady-state frequency amplitudes of musical instruments as coupled subsystems," in *Proceedings of International Symposium on Music Acoustics (ISMA)*, 2010.

[13] S. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, Ph.D. thesis, Stanford University, 1999.

[14] X. Serra, *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*, Ph.D. thesis, Stanford University, 1989.

[15] C. Duxbury, M. E. Davies, and M. B. Sandler, "Separation on transient information in musical audio using multiresolution analysis techniques," in *Proceedings of DAFx*, 2001.

[16] Laurent Daudet, "Transients modelling by pruned wavelet trees," in *Proc. ICMC*, 2011.

[17] Florian Keiler, Can Karadogan, Udo Zölzer, and Albrecht Schneider, "Analysis of transient musical sounds by autoregressive modeling," in *Proceedings of DAFx*, 2003.

[18] Benedict Tan and Deep Sen, "The use of the attack transient envelope in instrument recognition," in *Proc. Eleventh Australasian International Conference on Speech Science and Technology*, 2006.

[19] M.W. Macon, A. McCree, Wai-Ming Lai, and V. Viswanathan, "Efficient analysis/synthesis of percussion musical instrument sounds using an all-pole model," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1998.

[20] J. Laroche and J.-L. Meillier, "Multichannel excitation/filter modeling of percussive sounds with application to the piano," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 329–344, 1994.

[21] E. Ravelli, G. Richard, and L. Daudet, "Extending fine-grain scalable audio coding to very low bitrates using overcomplete dictionaries," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[22] R. Gribonval, P. Depalle, X. Rodet, E. Bacry, and S Mallat, "Sound signals decomposition using a high resolution mathing pursuit," in *Proceedings of the International Computer Music Conference*, 1996.

[23] B. Sturm, J. J. Shynk, L. Daudet, and C. Roads, "Dark energy in sparse atomic decompositions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 671–676, 2008.

[24] G. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, "An extension of the adaptive quasi-harmonic model," in *Proceedings of the International Conference on Audio, Speech and Signal Processing*, 2012.

[25] G. Kafentzis, O. Rosec, and Y. Stylianou, "On the modeling of voiceless stop sounds of speech using adaptive quasi-harmonic models," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2012.

[26] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the properties of a time-varying quasi-harmonic model of speech," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2008.

[27] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 2, pp. 290–300, 2011.