EVALUATING HOW WELL FILTERED WHITE NOISE MODELS THE RESIDUAL FROM SINUSOIDAL MODELING OF MUSICAL INSTRUMENT SOUNDS

Marcelo Caetano[1*], *George Kafentzis*[2,3], *Gilles Degottex*[1,2], *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,mouchtar,styliano}@ics.forth.gr, {kafentz,degottex}@csd.uoc.gr

ABSTRACT

Nowadays, sinusoidal modeling commonly includes a residual obtained by the subtraction of the sinusoidal model from the original sound. This residual signal is often further modeled as filtered white noise. In this work, we evaluate how well filtered white noise models the residual from sinusoidal modeling of musical instrument sounds for several sinusoidal algorithms. We compare how well each sinusoidal model captures the oscillatory behavior of the partials by looking into how "noisy" their residuals are. We performed a listening test to evaluate the perceptual similarity between the original residual and the modeled counterpart. Then we further investigate whether the result of the listening test can be explained by the fine structure of the residual magnitude spectrum. The results presented here have the potential to subsidize improvements on residual modeling.

Index Terms— Sinusoidal Analysis, Residual Modeling, Linear Prediction, Inverse Filter, Spectral Whitening.

1. INTRODUCTION

Sinusoidal modeling stands out among the models used to represent [1, 2, 3, 4] and transform musical instrument sounds [5, 6, 7] due to the fidelity and flexibility of the representation. In essence, sinusoidal analysis models each partial with a time-varying sinusoid, capturing temporal variations in amplitude, frequency and phase (the parameters of the model). Sinusoidal modeling represents musical instrument sounds well because most musical instruments are designed to present very clear modes of vibration. However, there is noise present in virtually all musical instrument sounds, such as breathing noise in woodwinds or mechanical noise like the hammer striking the piano strings.

There have been improvements in sinusoidal modeling to address issues such as partial tracking [2, 3, 4], transient modeling [8, 7], to augment the accuracy of parameter estimation as well as the temporal resolution by adapting partials' trajectories inside the analysis window [9, 10]. Nevertheless, the lack of noise is perceptually noticeable in the sinusoidal representation of musical instrument sounds [11, 12]. Serra [13] proposed to subtract the sinusoidal component (i.e., the result of sinusoidal analysis) from the original recording to estimate a "residual component". This residual is, by definition, whatever is left from sinusoidal modeling, and therefore, commonly assumed to be noise not captured by the sinusoidal model (usually because sinusoids are not a compact representation of noise). Considerably less effort has been made in residual modeling. It has become standard practice [13, 11, 12] to model the residual component by filtering white noise with a time-varying filter that emulates the spectral characteristics of the residual signal. Naturally, there are different ways to model the spectral distribution of energy of the residual component. The basic assumption is that the residual signal does not contain perceptually relevant information in the phase spectrum, only in magnitude. Therefore, "psychoacoustic" filter banks are usually found in residual modeling [7, 11]. Goodwin [11] uses the short-time energy in equivalent rectangular bands (ERBs) of the magnitude spectrum for both the analysis and synthesis stage, and justifies stating that the ear is insensitive to energy distributions within each ERB. Levine [7] uses Bark bands instead. Resynthesis commonly uses a piece-wise constant spectrum with magnitudes from the ERB (or Bark bands) energy and random phase. Goodwin remarks that temporal phase correlations can control the texture of the modeled residual, which has been studied further to synthesize environmental sounds (e.g., running water or crackling fire)[14]. Ding [12] proposes to use multi-pulse excitation linear prediction (MPLP) to keep phase coherence with the sinusoidal component.

There have been no formal investigations on the filtered white noise model for residual from sinusoidal modeling of musical instrument sounds. In this work we present a systematic evaluation of how well filtered white noise models the residual from sinusoidal modeling of musical instrument sounds for different sinusoidal modeling algorithms. Each algorithm captures oscillatory behavior differently and, consequently, leaves (perceptually) different residuals. We performed a subjective listening test to evaluate the perceptual similarity between filtered white noise and the residual of each sinusoidal algorithm. Then we use an objective measure of similarity to compare with the perceptual assessments. The next section briefly reviews the sinusoidal modeling algorithms used in this investigation. Next, we describe the framework used to decompose the musical instrument sounds into the blocks used in the evaluation, which is followed by a discussion and the conclusions and future perspectives.

2. SINUSOIDAL MODELING

Conceptually, traditional sinusoidal modeling supposes that the musical instrument sounds being modeled can be decomposed into quasi-harmonic oscillations and additive noise. In practice, the mu-

*This work is cofunded in part 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," and by the FP7-PEOPLE-IAPP "AVID-MODE" grant.

sical instrument sound $y(t)$ is separated into a sinusoidal component $y_s(t)$ plus a residual component $y_r(t)$, where $y_r(t)$ is obtained by subtraction of the purely sinusoidal component $y_s(t)$ from the original sound $y(t)$. The sinusoidal component is further represented as

$$y_s(t) = \left[\sum_{k=0}^{K} \alpha_k e^{j2\pi t f_k}\right] w(t) \quad (1)$$

where α_k and $\phi_k(t) = 2\pi t f_k$ are respectively the amplitude and phase of the k^{th} sinusoid inside the analysis window $w(t)$, and K is the number of sinusoids. The model assumes that the sinusoids describe stable partials of the sound so their parameters do not vary significantly inside the analysis window. Traditionally [13], the parameters of the model α_k and $\phi_k(t)$ are estimated for each frame of the short-time Fourier transform, limiting the temporal resolution of the model to that of the STFT. In this article, SM stands for a sinusoidal model that imposes no restrictions on the frequencies of the partials [13]. For most musical instrument sounds, a model where the sinusoids are harmonically related is a good approximation, giving rise to the harmonic model (HM) [15], which uses sinusoids whose frequencies are multiple integers k of a fundamental frequency f_0 as $\phi_k(t) = 2\pi t k f_0$.

There have been proposals to improve the temporal resolution of the sinusoidal model by adapting the estimation of the parameters of the sinusoids *inside* the analysis window, resulting in *adaptive* sinusoidal models. In particular, the adaptive harmonic model (aHM) [9] used in this work modulates the frequency of each sinusoid inside the analysis window upon resynthesis. Recently, the extended adaptive quasi-harmonic model (eaQHM) was developed [10]. The eaQHM algorithm adapts both the amplitudes and frequencies of the sinusoidal partials inside the analysis window, therefore it can be considered a full AM/FM model, as shown below

$$y_s(t) = \left[\sum_{k=0}^{K} \alpha_k(t)\, e^{j\phi_k(t)}\right] w(t), \quad (2)$$

where $\alpha_k(t)$ denotes the time-varying amplitude and $\phi_k(t)$ denotes the instantaneous phase function of the k^{th} component inside the analysis window $w(t)$. Table 1 summarizes the temporal representation of frequencies for the analysis and synthesis stages for the sinusoidal algorithms used.

2.1. Residual Modeling

The residual component $y_r(t)$ is modeled as

$$\hat{y}_r(t) = \int_0^t a(t-\tau)\, u(\tau)\, d\tau \quad (3)$$

where $\hat{y}_r(t)$ is the modeled residual component, $u(\tau)$ is white noise and $a(t,\tau)$ is the response of a time-varying filter. Serra [13] wrote that "a stochastic, or noise, signal is fully described by its power spectral density which gives the expected signal power versus frequency. When a signal is assumed stochastic, it is not necessary to preserve either the instantaneous phase or the exact magnitude details of individual FFT frames," justifying the assumption that the residual component can be modeled as filtered white noise. There have been different proposals to estimate the filter $a(\tau)$ [13, 11, 12]. In this work, we estimate the spectral envelope of each frame of the STFT of the residual component $y_r(t)$ using linear prediction (LPC) [16] and use it as the time-varying filter coefficients, as has been previously proposed for speech [15]. LPC is adequate for spectral envelope estimation of $y_r(t)$ because it tends to follow the average energy of noisy spectra rather than the peaks. Using eq. (3) the model suposes that if we inverse filter $y_r(t)$, we should obtain white noise (a signal with flat magnitude and no temporal phase coherence or random phase). In this work, we investigate if filtered white noise is perceptually close to the original residual signals with a listening test and further investigate if the inverse filtered residual component presents the characteristics of white noise with an objective measure based on the autocorrelation function.

Table 1: Comparison of representations of frequency components for the analysis and synthesis stages of the sinusoidal algorithms used.

	Analysis	Synthesis
SM	stationary	stationary (OLA)
HM	stationary	Splines
aHM	adaptive	Splines
eaQHM	adaptive	Splines

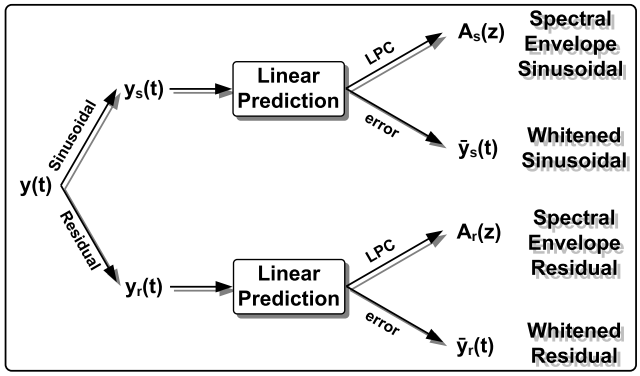

Figure 1: Illustration of the signal decomposition.

3. EXPERIMENTAL FRAMEWORK

Figure 1 illustrates the steps of the experimental framework. Each musical instrument sound $y(t)$ is decomposed into sinusoidal $y_s(t)$ and residual $y_r(t)$ using SM, HM, aHM, and eaQHM. Each component, $y_s(t)$ and $y_r(t)$, is modeled with linear prediction, resulting in a time-varying spectral envelope $A_s(z)$ and $A_r(z)$ and an inverse filtered (whitened) signal $\bar{y}_s(t)$ and $\bar{y}_r(t)$, which are the prediction errors [16]. In the listening test, we use white noise filtered with $A_r(z)$. The objective similarity measure compares $\bar{y}_r(t)$ with $\bar{y}_s(t)$ and $u(t)$.

Table 2 lists the 14 musical instruments used [1]. The pitch of all sounds is $C3 \simeq 131$ Hz, the *dynamics* is *forte*, and the duration is less than 2s. All sinusoidal algorithms used a window size equivalent to 3 times the period of the fundamental frequency $f_0 \simeq 131$ Hz, 50% overlap, and size of the FFT 4 times the window size. The linear prediction order used was 50 for both $y_s(t)$ and $y_r(t)$ to avoid smearing possible oscillatory energy left in $y_r(t)$ (missed by the sinusoidal model).

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

4. EVALUATION

The evaluation consists of a listening test and an objective measure based on the autocorrelation function. However, firstly we estimate the residual energy to compare how well each sinusoidal algorithm models the musical instrument sounds. The less residual energy, the better the algorithm captured the oscillatory behavior. The signal to reconstruction error ratio (SRER) shown in eq. (4) measures the ratio between the total energy and the energy in the residual component $y_r(t)$. The higher the ratio, the less residual energy there is in $y_r(t)$.

$$\text{SRER (dB)} = 20\log_{10}\frac{\text{RMS}\left[y(t)\right]}{\text{RMS}\left[y_r(t)\right]} \tag{4}$$

where $y(t)$ is the original signal and $y_r(t)$ is the residual component. Table 3 shows the average SRER in dB across musical instrument sounds for each method, revealing that eaQHM has a higher SRER than all other methods by roughly 15 dB.

4.1. Listening Test

The purpose of the listening test is to evaluate the perceptual similarity between the residual signal $y_r(t)$ and its filtered-white-noise counterpart $\hat{y}_r(t)$ for the the 14 musical instrument sounds listed in Table 2 modeled with the four sinusoidal algorithms shown in Table 1. For each participant, the listening test presented a subset of 16 pairs of sounds corresponding to $y_r(t)$ and $\hat{y}_r(t)$ from 4 musical instruments (times 4 algorithms) in random order to minimize cross comparison among methods. All sounds were normalized at -16dB RMS. The listener is instructed to listen to each pair as many times as they want and rate their perceptual similarity in a scale from 1 to 5 labeled with the terms 1) *Very different*, 2) *Different*, 3) *Fairly similar*, 4) *Very similar*, 5) *Identical*. The test can be found at `http://gillesdegottex.eu/ExCaetano2013simil`. Figure 2 shows the result for 51 participants aged between 22 and 67, depicting the mean opinion score (MOS) and 95% confidence interval. In average, eaQHM results in a residual signal that was considered between *fairly similar* and *very similar* to its filtered white noise counterpart. The other 3 algorithms (SM, HM, and aHM) produced residuals whose filtered white noise counterparts were considered practically *different*.

4.2. Objective Measure

The result of the listening test indicates that, in general, filtered white noise was not considered a perceptually similar representation of $y_r(t)$. However, the listening test gives no further evidence to help explain why. Ideally, we would like to identify what remains in the residual signal that departs from the conceptual filtered-white-noise hypothesis. In the listening test, the perceptual effect of the LPC spectral envelope of $y_r(t)$ is present in $\hat{y}_r(t)$. Thus we assume that the differences lie elsewhere, in the spectral fine structure or in the phase spectrum. To evaluate the importance of the fine structure between $y_r(t)$ and $\hat{y}_r(t)$, we compare the whitened residual component $\bar{y}_r(t)$ with the whitened sinusoidal component $\bar{y}_s(t)$ and with the model (i.e., white noise $u(t)$) with an objective similarity measure. We use the autocorrelation functions, shown in (5), which should provide a unique representation of both the white noise (zero except at zero lag) and the sinusoidal component (peaks at multiple integers of the fundamental frequency).

$$R(i) = \sum_{n=0}^{N-1-i} y(n)\, y(n-i) \tag{5}$$

The similarity measure is then the dot product between the autocorrelation functions, given by $\cos\left(\Theta\left\{\bar{y}_r, u\right\}\right) = R_{\bar{y}_r}(i) \cdot R_u(i)$ and $\cos\left(\Theta\left\{\bar{y}_r, \bar{y}_s\right\}\right) = R_{\bar{y}_r}(i) \cdot R_{\bar{y}_s}(i)$. The dot (or inner) product can be interpreted as the projection of $R_{\bar{y}_r}(i)$ onto $R_{\bar{y}_s}(i)$ and $R_u(i)$. Thus Θ is the angle between the autocorrelation functions interpreted as vectors, and it varies from 0 (identical) to 90° (orthogonal). Table 4 shows the average of these values across all musical instruments to allow comparison per method. Following Fig. 2, we expected eaQHM to give a significantly smaller $\Theta\left\{\bar{y}_r, u\right\}$ and larger $\Theta\left\{\bar{y}_r, \bar{y}_s\right\}$.

Table 2: Musical instrument sounds used in the listening test.

Strings	**Brass**	**Woodwinds**
Double Bass	Bass Trombone	Bass Clarinet
Cello	Bass Trumpet	Bassoon
Viola	Cimbasso	Clarinet Bb
	Contrabass Tuba	English Horn
	Tenor Trombone	
	Tuba	
	Wagner Tuba	

Table 3: Average Signal to Reconstruction Error Ratio (SRER) across musical instrument sounds.

SRER (dB)

SM	HM	aHM	eaQHM
33.86	34.84	36.53	50.62

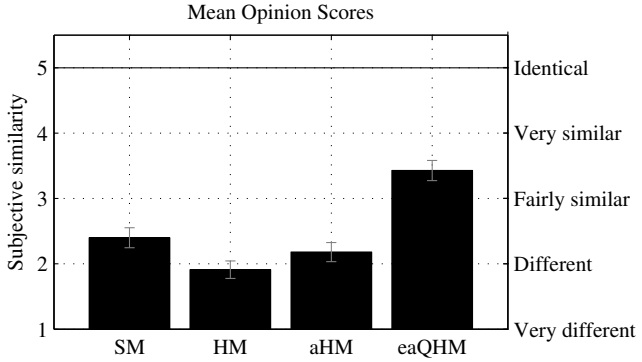

Figure 2: Result of the listening test. The figure shows the mean opinion score (MOS) and 95 % confidence interval for the four sinusoidal models tested.

Table 4: Average angle in degrees across musical instrument sounds for each algorithm.

	SM	**HM**	**aHM**	**eaQHM**
$\Theta\left\{\bar{y}_r, u\right\}$	46.11°	51.63°	49.83°	50.95°
$\Theta\left\{\bar{y}_r, \bar{y}_s\right\}$	61.46°	67.25°	68.85°	67.48°

5. DISCUSSION

Each sinusoidal modeling algorithm resulted in a different perceptual similarity, revealing that different algorithms leave different undesired information in the residual signal y_r. Therefore we suspect that there might be some oscillatory behavior left in y_r. In other words, some sinusoidal modeling algorithms fail to capture all oscillatory energy such as frequency modulations or transients. The models that use slowly varying sinusoids (stable oscillations) plus additive noise might oversimplify the complexity of musical sounds. It has already been remarked [7] that *sinusoids* plus *noise* plus *transients* might be a more realistic representation for musical instrument sounds. Transients are characteristically present mostly during the attack, but there is no indication that the participants used the attack as perceptual cue. The listening test shows that the AM/FM modeling of eaQHM captures most oscillatory energy, including transients.

On the other hand, Table 4 reveals no significant difference across algorithms. The angles Θ do indicate that $\bar{y}_r$ is closer to u (white noise) than to $\bar{y}_s$ (sinusoidal) for all algorithms. But the similarities measured by Θ do no explain the results of the listening test. Our interpretation of this result is that the perceptual differences found in the listening test cannot be explained by fine spectral structure, rather, by phase coherence or transients.

Interestingly, one of the participants of the listening test remarked that, for each pair, one of them always sounded *brighter*. Indeed, $\hat{y}_r$ has more energy in high frequencies because pure white noise has a flat spectrum where energy is not equal per octave (let alone per ERB or Bark band). A possible course of investigation would be to use different types of noise (prior to applying the time-varying spectral envelope) to correctly balance the spectral energy, such as *pink* noise.

6. CONCLUSIONS AND FUTURE PERSPECTIVES

We presented a systematic investigation of the filtered white noise model for the residual from sinusoidal modeling of musical instrument sounds. Four different sinusoidal modeling algorithms were evaluated. We conducted a listening test and we developed an objective measure of spectral similarity. The listening test assessed the perceptual similarity between filtered white noise and the residual component for each sinusoidal algorithm. The results indicate that, in general, filtered white noise was considered *different* from the residual component. However, we determined that eaQHM leaves a residual that is *fairly similar* to the filtered white noise counterpart. The objective measure compared the residual with both the sinusoidal component and their modeled counterpart across algorithms using the autocorrelation functions. The objective evaluation aimed to investigate the reason for the result of the listening test, trying to indicate whether there was "sinusoidal" energy left in the poorly modeled residuals, for example. The objective similarity measure did not indicate that the perceptual differences found can be explained by comparing spectral fine structure. However, the autocorrelation function only includes information from the power spectral density. Thus we suspect that the differences lie in the phase spectrum (possibly due to temporal phase coherence) or transients in the residual, confirming the conclusion of previous studies [11, 7].

Future work should focus on determining the reason for the difference between the conceptual model of filtered white noise and what current sinusoidal modeling algorithms fail to model. Perspectives include using "colored" noise to correct the high-frequency energy content perceived as brightness (or some other more sophisticated psychoacoustic model). Further investigation on the temporal phase coherence should develop a measure for analysis and comparison with the sinusoidal component. Attack transients might account for some of the perceptual difference we found for most sinusoidal algorithms.

7. REFERENCES

[1] X. Serra and J. 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.

[2] L. Nunes, R. Merched, and L. W. P. Biscainho, "Recursive least-squares estimation of the evolution of partials in sinusoidal analysis," in *Proc. ICASSP*, 2007.

[3] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1625–1634, 2007.

[4] P. Depalle, G. Garcia, and X. Rodet, "Tracking of partials for additive sound synthesis using hidden markov models," in *Proc. ICASSP*, 1993.

[5] X. Serra and J. Bonada, "Sound transformations based on the sms high level attributes," in *Proc. DAFX*, 1998.

[6] X. Amatriain, J. Bonada, A. Loscos, and X. Serra, "Spectral processing," in *DAFX - Digital Audio Effects*, U. Zolzer, Ed. John Wiley and Sons, 2002, ch. 10, pp. 373–438.

[7] S. N. Levine and J. O. Smith Iii, "A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modications," in *Proc. AES Convention*, 1998.

[8] T. S. Verma and T. H. Y. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music Journal*, vol. 24, no. 2, pp. 47–59, 2000.

[9] G. Degottex and Y. Stylianou, "A Full-Band Adaptive Harmonic Representation of Speech," in *Proc. Interspeech*. ISCA, 2012.

[10] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, "An Extension of the Adaptive Quasi-Harmonic Model," in *Proc. ICASSP*, 2012.

[11] M. Goodwin, "Residual modeling in music analysis-synthesis," in *Proc. ICASSP*, 1996, pp. 1005–1008.

[12] Y. Ding and X. Qian, "Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (quasar) signal model," *J. Audio Eng. Soc*, vol. 45, no. 7/8, pp. 571–584, 1997.

[13] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. D. Poli, Eds. Swets & Zeitlinger, 1997.

[14] M. Athineos and D. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," in *Proc. WASPAA*, 2003.

[15] Y. Stylianou, "Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification," Ph.D. dissertation, TelecomParis, France, 1996.

[16] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, 1975.