Aalborg Universitet

# Stochastic analysis/synthesis using sinusoidal atoms

Jensen, Kristoffer

*Published in:*
Proceedings of the Frontiers of Research on Speech and Music

*Publication date:*
2008

*Document Version*
Early version, also known as pre-print

Link to publication from Aalborg University

*Citation for published version (APA):*
Jensen, K. (2008). Stochastic analysis/synthesis using sinusoidal atoms. In A. Datta (Ed.), *Proceedings of the Frontiers of Research on Speech and Music* (pp. 340-347). ITC Sangeet Research Academy, Kolkata, India.

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

**Take down policy**
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: April 25, 2024

# STOCHASTIC ANALYSIS/SYNTHESIS USING SINUSOIDAL ATOMS

Kristoffer Jensen

Aalborg University Esbjerg
Niels Bohr Vej 8, 6700 Esbjerg, Denmark
krist@aaue.dk

***Abstract.*** *This work proposes a method for re-synthesizing music for use in perceptual experiments regarding structural changes and in music creation. Atoms are estimated from music audio, modelled in a stochastic model, and re-synthesized from the model parameters. The atoms are found by splitting sinusoids into short segments, and modelled into amplitude and envelope shape, frequency, time and duration. A simple model for creating envelopes with percussive, sustained or crescendo shape is presented. Single variable and joint probability density functions are created from the atom parameters and used to re-create sounds with the same distribution of the atoms parameters. A novel method for visualization music, the musigram, permits a better understanding of the re-synthesized sounds.*

## 1 Introduction

Sinusoidal Analysis/Synthesis (A/S) is the task of estimating time-varying frequencies and amplitudes of sinusoids from a sound, which can, after appropriate processing, be resynthesized with little loss of quality. In this work, the processing consists of randomizing the time, duration, frequency and amplitudes of the atoms, using probability density functions (pdf) obtained form the original sounds, If appropriate pdfs are used, much of the original sound is recreated, notwithstanding the stochastic nature of the processing. This process has been applied to instrument sounds and longer music excerpts.

In this work, the goal is not so much to recreate the sounds without loss of quality, but instead gather a better understanding of the dependencies of the common features of the sound (time, amplitude, frequency, …). Such an understanding is helpful in many situations regarding perception and processing of sounds. Furthermore, this work is undertaken as the next step in a series of models regarding noise created from sinusoidal atoms. Indeed, in Jensen (2005a), the first step was taken, by showing that atoms randomly spread in time and frequency could recreate all kind of noises, from *Geiger*-like clicks to cymbal-like almost voiced sounds. Furthermore, tone was re-inserted through periodic *pdfs*, either on the frequency, or on the time, rendering insect like or eerie, almost-natural periodic sounds. Finally, tone was created by repeating the noise periodically, which rendered harsh periodic tone with uncertain pitch. This work was enhanced in (Jensen 2005b), which adjusted the density and amplitude of the atoms to be perceptually white and linear.

Another motivation for this work is to create stimuli for experiments regarding the perception of structure change. Kühl & Jensen 2008 suggested through understanding of the temporal cognition and analysis of music that structural change should occur at approximately 30 seconds interval. It is a hypothesis that the work presented here could produce music with or without structural changes, dependent on which pdfs are used.

This paper is structured as follows; in chapter 2, the atoms are extracted from audio, and modelled. In chapter 3, the parameters from the atoms are statistically modelled, and in chapter 4, sounds are re-synthesized using the statistical data.

# 2 Parameter estimation

The element used in the stochastic resynthesis is called an atom, and it is obtained from sinusoidal analysis. This involves estimating the time-varying frequency and amplitude of pure sinusoids. There exist a multitude of methods for estimating the parameters of the sinusoids, most are based on the Fourier transform. In such a program, the sound to be analysis is split into short segments, and the peaks of the Fourier transform are attached to the previous partials tracks to form longer sinusoids (McAulay & Quatieri 1986). The sound can be recreated with good accuracy using the model in eq. 1, but better results are obtained, if the relative phase of each sinusoid is retained (Andersen & Jensen 2004).

$$s(t) = \sum_{k=1}^{N} a_k(t) \cdot \sin(2\pi \int_{\tau=0}^{t} f_k(\tau) d\tau) . \qquad (1)$$

The atoms are found by detecting zeros in the amplitude, and they are modelled by the mean amplitude, maximum frequency, duration and time, and shape of the amplitude envelope.

## 2.1 Sinusoidal analysis

The analysis is done here using the perceptual criteria hearing threshold and masking (Zwicker & Fastl 1990) to only retain audible sinusoids. The audible sinusoids are estimated for each time step, and attached to previous partials tracks, if the frequency distance is small enough. The frequency resolution is improved by performing second order interpolation using three FFT bins for each peak. If a partial track is not attached to a current partial, the track is ended, and it is considered one atom. The result creates a few hundred atoms for each second of sound, corresponding to the harmonic and non-harmonic voiced part, and the unvoiced part of the sound.

The resulting atoms are modelled by their mean frequency, maximum amplitude, duration and start time. In addition, the time centroid is calculated, in order to permit the resynthesis of percussive or sustained atoms.

## 2.2 Resolution and scaling

In order to respect the human hearing perception, it is necessary to analyze the sounds with a fine time and frequency resolution. Unfortunately, a good time resolution (using a short time window) corresponds to a bad frequency resolution, and vice-verse. Therefore, a less than perfect compromise is usually found. In this work, a time resolution of 10 msec is used. While this gives a good enough re-synthesis sound quality for most sounds, the resolution is too fine for the statistical histograms. Therefore, the frequency is furthermore transformed into log frequency, as a crude model of the frequency discrimination of the ear, and into chromas,

$$c_k = \log_2(12 f_k) \;\; modulo \;\; 12 \qquad (2)$$

in order to capture to common 12 notes per octave found in most popular music. Similarly, the amplitudes are converted into dB, which is closer to the human sensation of loudness.

## 2.3 Atom amplitude

An analysis of the atom envelope reveals rather static frequencies, and amplitudes with zero at the extremities and positive values in between. After resampling and normalization, the amplitudes are further analyzed; while all values between zero and one exists for all time steps, except the first and last, the mean at each time-step reveals rounded increasing and decreasing envelope with maximum at half-distance. Further analysis of the mean resampled and normalized amplitude envelopes for music and instrument sounds reveals that this type of curve is found in all situations, but it is slightly different for each sound source, and the time centroid is always close to one half. The isolated instrument sounds analyzed has lower time centroid and more peaked envelopes, compared to the music sounds.

The atoms with a low time centroid has a percussive sound, while those with a high time centroid are crescendo-like. Therefore, the time centroid is potentially important. It is modelled by the square root of a Bartlett-Hamming window, using a two-slope linear time to model the skewness of the envelope.

$$\mathrm{w} = \sqrt{0.62 - 0.48 \cdot |t| + 0.38 \cdot \cos(2pt)} \tag{3}$$

where $t$, symmetric around zero, has slope $s$ until $N$, and slope $1/s$ until $L$.

$$\mathrm{N} = \mathrm{s(L + s - 1)}/s^2 + 1 \tag{4}$$

and

$$\mathrm{s} = \frac{1 + \sqrt{1 + 4\mathrm{L}(\mathrm{t}_0 - 1) + 4\mathrm{t}_0(1 - \mathrm{t}_0)}}{2(\mathrm{t}_0 - 1)}. \tag{5}$$

$L$ is the length of the envelope, and $t_0$ is the position of the peak of the envelope. Using eq. 3-5, it is now possible to create atoms with varying skewness, rendering percussive ($t_0 < L/2$) sustained ($t_0 = L/2$), or crescendo ($t_0 > L/2$) shape. Examples of such atom envelopes are shown in figure 2 together with the means of many time and amplitude normalized envelopes with low, medium and high time centroid values. While, of course, the individual atom envelopes have any and all possible shapes, this model at least fits the mean normalized envelopes well.

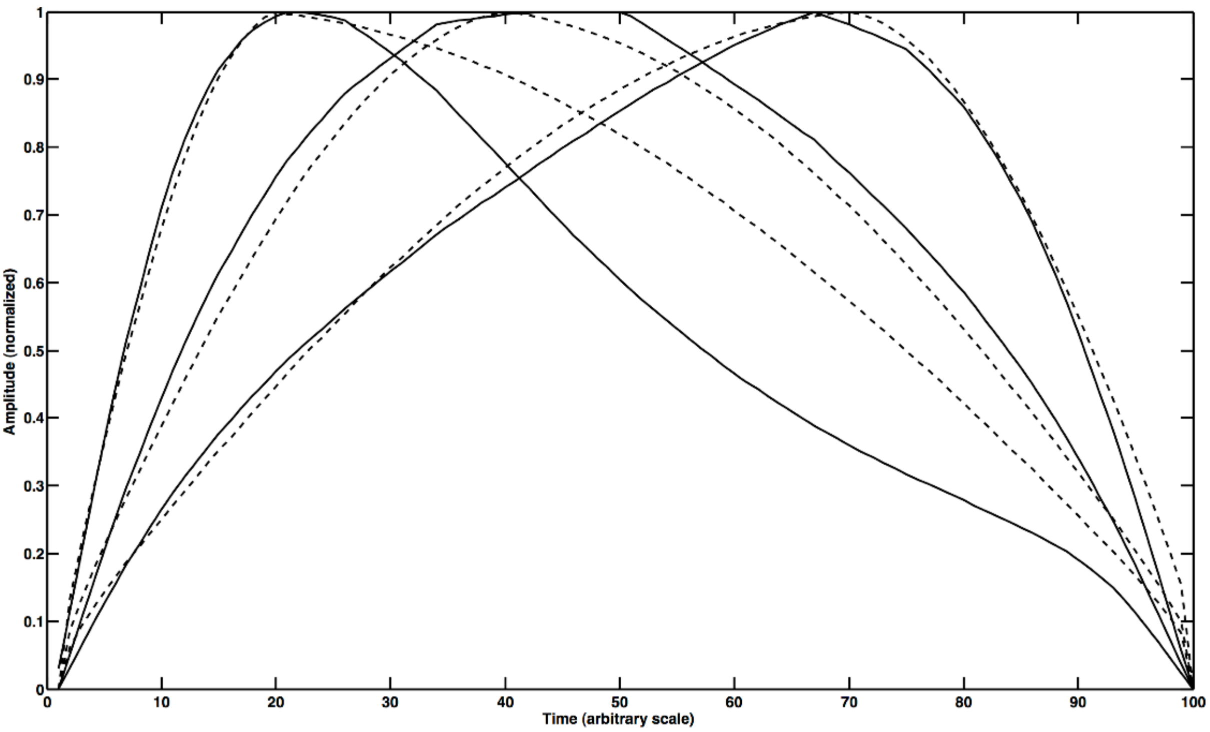


**Figure 1. Means of normalized amplitude envelopes for low, medium and high time centroid (solid) with modelled envelopes from eqs. 3-5 (dashed).**

## 3 Statistical modelling

The atoms are now used for statistics, by counting the number of occurrences (histogram) of each frequency, amplitude, duration and time step. As the parameters are not independent, the joint probabilities of several parameters are also calculated. This is to be used in the situation where the parameters are not independent, which they are more often than not, in particular between the log frequency and amplitude, between the chroma frequency and the time and between the log frequency and the durations.

### 3.1 Histograms

In a first step, the number of occurrences of the discrete values of log frequency, chroma, amplitude in dB, duration and time is counted for a song. As an example, the histograms of a 20 seconds excerpt of a song of Suchitra Mitra[1] are shown in figure 2. This being an old recording, only low frequencies are found, below 5 kHz. It has approximately 30 dB dynamic range, and the atoms found are all below 300 ms. Finally, as expected, it has atoms spread out across the full 20 seconds, intermittently, corresponding to the attacks and other lively moments of the song.

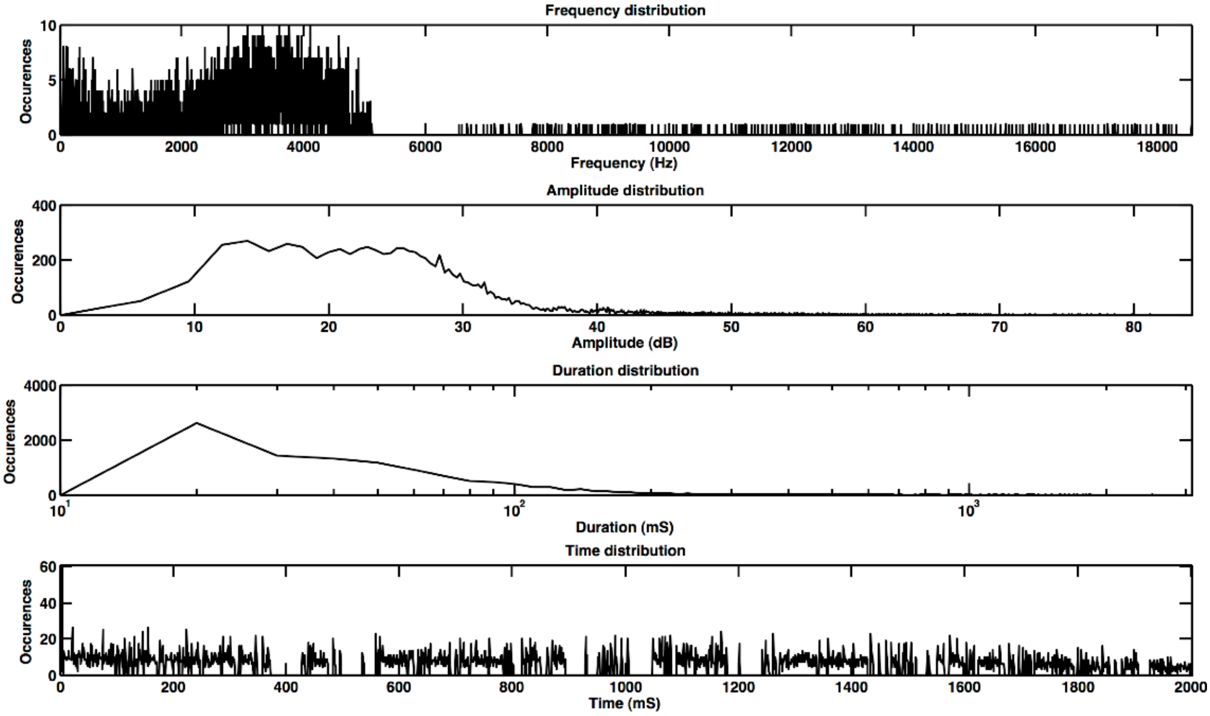


**Figure 2. Histograms of the frequencies, amplitudes, durations and time onsets from a Bengali song[1].**

In order to be usable in the following, the histograms are normalized into probability density functions (pdf).

### 3.2 Joint probability

While it is possible to use the results in the previous chapter to resynthesize a sound, this would not be very interesting, as music is not static. Therefore, in order to obtain perceptually convincing results, several observations about the dependencies between the estimated parameters are made. These dependencies are then to be integrated into the resynthesis model in order to improve the quality.

It is easily found that the amplitudes are lower in high frequencies, as a general rule. The duration dependencies are not easily observed at this point, but they are mainly dependent on the frequencies. The time occurrences are dependent on the chroma frequency, to take into account melody, chord and tonality changes, for instance.

[1] Suchitra Mitra, Aaguner Parashmoni Chanao Pranay.

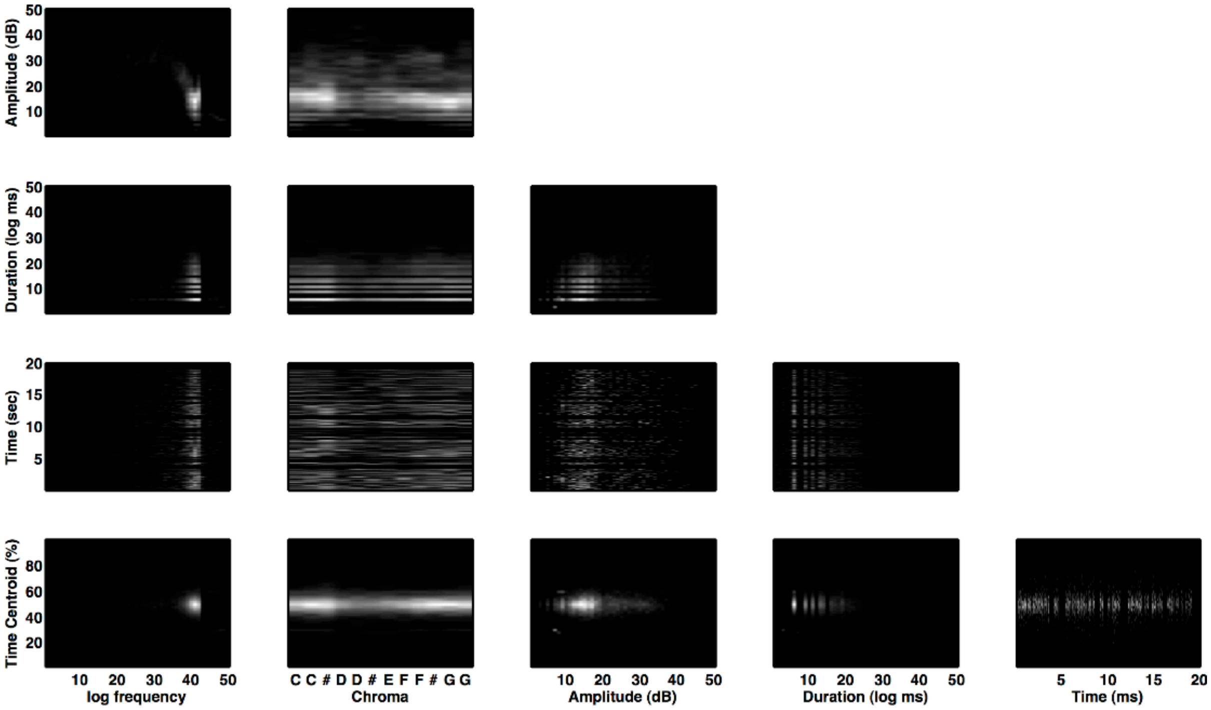
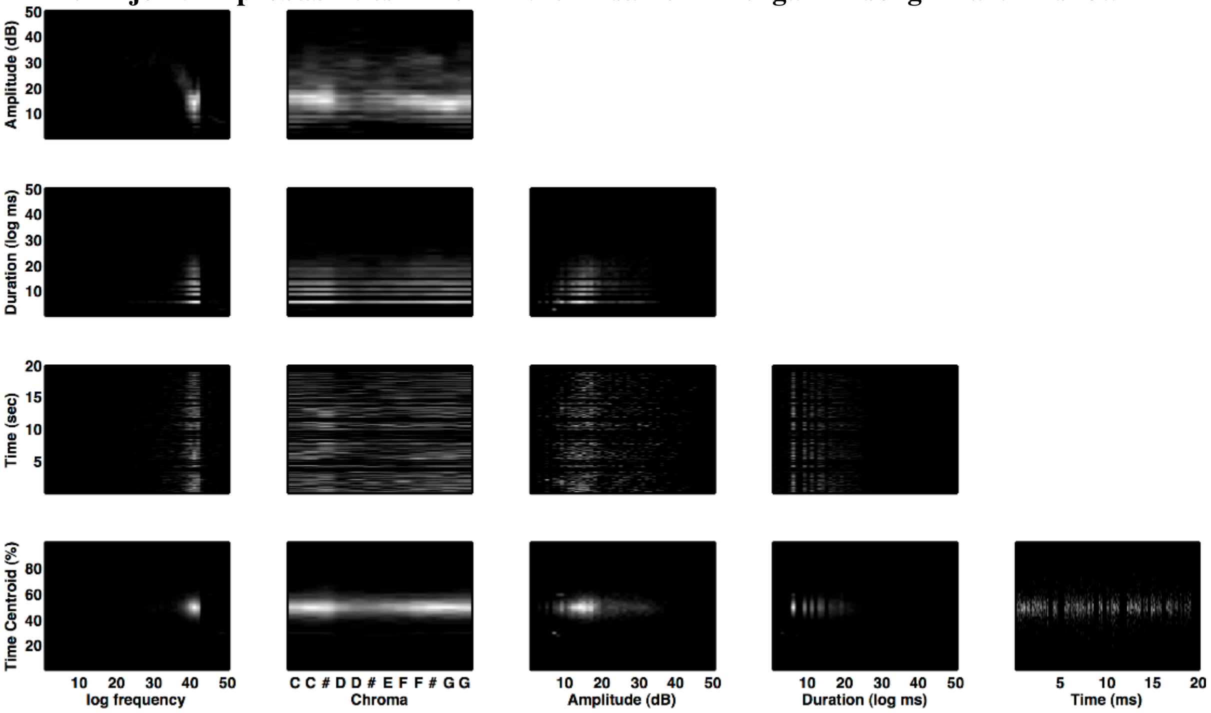


**Figure 3. Joint probabilities for frequency, chroma, amplitude, duration, time, and time centroid for an excerpt of a Bengali song.**



**The joint probabilities for the same Bengali song are shown in** figure 3. The interesting joint probabilities are then the dependent ones, such as the log frequency/amplitude joint probability, in which it is shown that low frequency atoms are louder. Other interesting joint probabilities are the log frequency and amplitude to duration, the chroma to time, and most of the time centroid ones, in particular the amplitude/time centroid joint probability.

# 4 Resynthesis

When resynthesizing the sound, a given (identical to the original sound) number of atoms are created, one by one. The parameters of the atoms are found from the (joint) probabilities density functions, using the inverse image method.

## 4.1 Which Joint probability?

In the case of joint probabilities, one of the random variables has to come first. This means, that one of the variables of each atom need to be created from the single variable pdf. Perhaps the most important parameter of music sounds is the frequency, and it is therefore obtained first by the inverse image method (Gray & Davisson 2004) from the frequency pdf. The obtained frequency is then transformed into log frequency and chroma.

For the other variables, the observations in section 3.2 are used and the joint pdfs are now used to obtain the amplitude (from the log frequency), duration (also from the log frequency), time (from the chroma frequency) and time centroid (from the amplitude) of the current atom.

This is repeated for all atoms that are subsequently added to the final sound output. The frequency of the atom is static, while the amplitude envelope is created utilizing eqs. 3-5. Experiments with the different joint pdfs indicates that this is a robust method for re-synthesizing the music.

## 4.2 Visualization of results

On a general term, it is possible to recognize the music or instrument sound that is recreated using the method described here. While the sounds recreated certainly are far from the original sounds, the main problem seems to be the lack of coherence in the partials, combined with ‘stray’ partials that are excessively long. These artefacts that stems from the stochastic nature of the synthesis, seems hard to avoid using the current 2D joint pdfs.

In order to assess the quality of the resynthesis, it has been subject to a number of visualization methods. These methods used here, dubbed *musigrams*, consists of visualization tools to illustrate the evolution of rhythm, timbre and chroma (*rhythmogram* (Jensen 2005c), *timbregram* and *chromagram* (Jensen 2006), respectively). The *rhythmogram* is created by visualizing the autocorrelation of the perceptually weighted spectral flux of overlapping blocks. It has rhythm interval on the y-axis, and it renders easily visualized information about the evolution of rhythm and tempo along the time. The *timbregram* and *chromagram* are both found from Gaussian time-weighted spectrograms, the *timbregram* from the PLP front-end (Hermanski 1990), and the *chromagram* from the chroma (eq. 2). The *timbregram* visualizes the timbre over time, with *bark* frequency on the y-axis, and the *chromagram* visualize the evolution of chroma over time, with chroma on the y-axis.

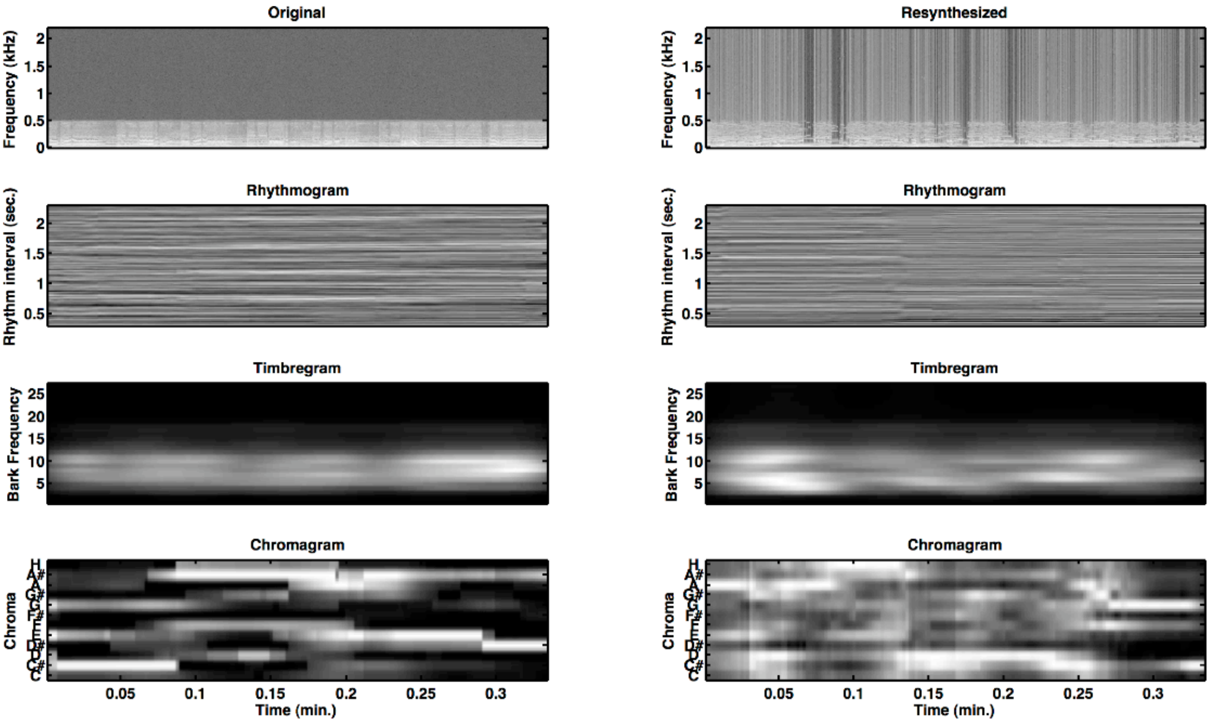


**Figure 4. Musigram of 20 seconds excerpt of a Bengali song. Spectrogram (top), *rhythmogram*, *timbregram* and *chromagram* (bottom). Original (left), resynthesized (right). White is strong, black weak.**

The *musigram* of the same Bengali song are shown in figure 4.

The main problem detected when listening to the resynthesized music, is the lack a synchronicity that prevents *fusion* between the overtones. This particular problem is not visible in the *musigrams*. Otherwise, the spectrogram (top) shows that the time/frequency content is retained, with the exception of perhaps too short atoms in the resynthesis that spread out too much in the high frequencies. This particular music being not very rhythmical, the *rhythmogram* (second from top) is not very informative. However, the rather stable, but not very enhanced rhythm is seemingly retained. The *timbregram* show that the general timbre is retained, but moved to be stronger in the start of the music. Finally, the *chromagram* is also similar, without recreating the note chromas exactly. The main problem with the resynthesis that can be understood from the analysis of the *musigrams* is that e.g. the amplitude is dependent on both time and frequency. As this is not modelled in the joint pdf, deviations from the original sound occurs, for instance in the chroma. An atom with a given chroma is placed on the right time, but without considerations of its amplitude. Other problems found with isolated sounds is that the strong, low frequency atoms, usually are longer than the high frequency atoms. This is not modelled in the current 2D joint pdf, as duration is obtained from the amplitude/duration joint pdf.

## 5 Conclusions

Common analysis and synthesis of music sounds involves modelling the sounds, and estimating the parameters of the model, and use them for resynthesis of the sound. In this work, the model is in two steps, first the sound is modelled into atoms, and then the atoms parameters are modelled in a stochastic model. The atoms are found from sinusoidal analysis using masking and hearing threshold, and modelled into amplitude, frequency, time and duration. The envelope of the atoms is modelled by the time centroid, and recreated using the square root of a Bartlett-Hamming window with a two-slope time function. This resembles the mean of the original normalized envelopes remarkable well. In the stochastic model, the parameters of the atom model are modelled in single feature or joint probability density functions. Analysis of the pdfs is used to determine which ones to use in the resynthesis. The music sounds are resynthesized by summing the atoms with parameters found by the inverse image method. The sounds are analyzed using the *musigrams*, a novel visualization tool that visualizes the evolution over time of rhythm, timbre and chroma. This shows that stochastic synthesis can bring out most of the identity of musical sounds and music. The main problems found are a lack of *fusion* of the overtones, and a multiple dependency found, i.e. the amplitude is dependent on both frequency and time.

Some of the uses planned of this work are; a better understanding of what is music can be found in listening experiments of recognition of music genre, instrument or other identities of the music, using sounds synthesized with the stochastic A/S. Similarly, this method can be used in creative work, to synthesize sounds that resembles, but are not identical to a given music. Finally, the statistic parameters of the sounds can be used in classification of music and musical sounds.

## References

Jensen, K. (2005a). *Atomic Noise*, Organised Sound, 10(1) pp. 75-81.

Jensen, K. (2005b) *Perceptual Atomic Noise*, Proceedings of the International Computer Music Conference, Barcelona, Spain, pp. 668-671.

McAulay, R. J., T. F. Quatieri (1984). *Speech analysis/synthesis based on a sinusoidal representation*, IEEE Trans. on Acoustics, Speech and Signal Proc., 34(4), pp. 744-754.

Kuhl, O. and K. Jensen (2008). *Retrieving Musical Form*, Lectures Notes in Computer Science. Accepted for publication.

Andersen, T. H. and K. Jensen (2004). *Importance and Representation of Phase in the Sinusoidal Model*, JAES, 52(11) pp. 1157-1169.

Zwicker E. and H. Fastl (1990). *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin, Heidelberg.

Robert M. Gray and L. D. Davisson (2004). *An Introduction to Statistical Signal Processing*, Cambridge University Press.

Jensen, K. (2005c). *A Causal Rhythm Grouping*. Lecture Notes in Computer Science, Volume 3310, pp. 83-95

Jensen, K. (2006). *Multiple scale music segmentation using rhythm, timbre and harmony*, EURASIP Journal on Applied Signal Processing, Special issue on Music Information Retrieval Based on Signal Processing.

Hermansky H. (1990). *Perceptual linear predictive (PLP) analysis of speech*, J. Acoust. Soc. Am., 87(4), pp. 1738-1752.

Sekey A. and B. A. Hanson (1984). *Improved 1-bark bandwidth auditory filter*. J. Acoust. Soc. Am., 75(6), pp. 151–168.