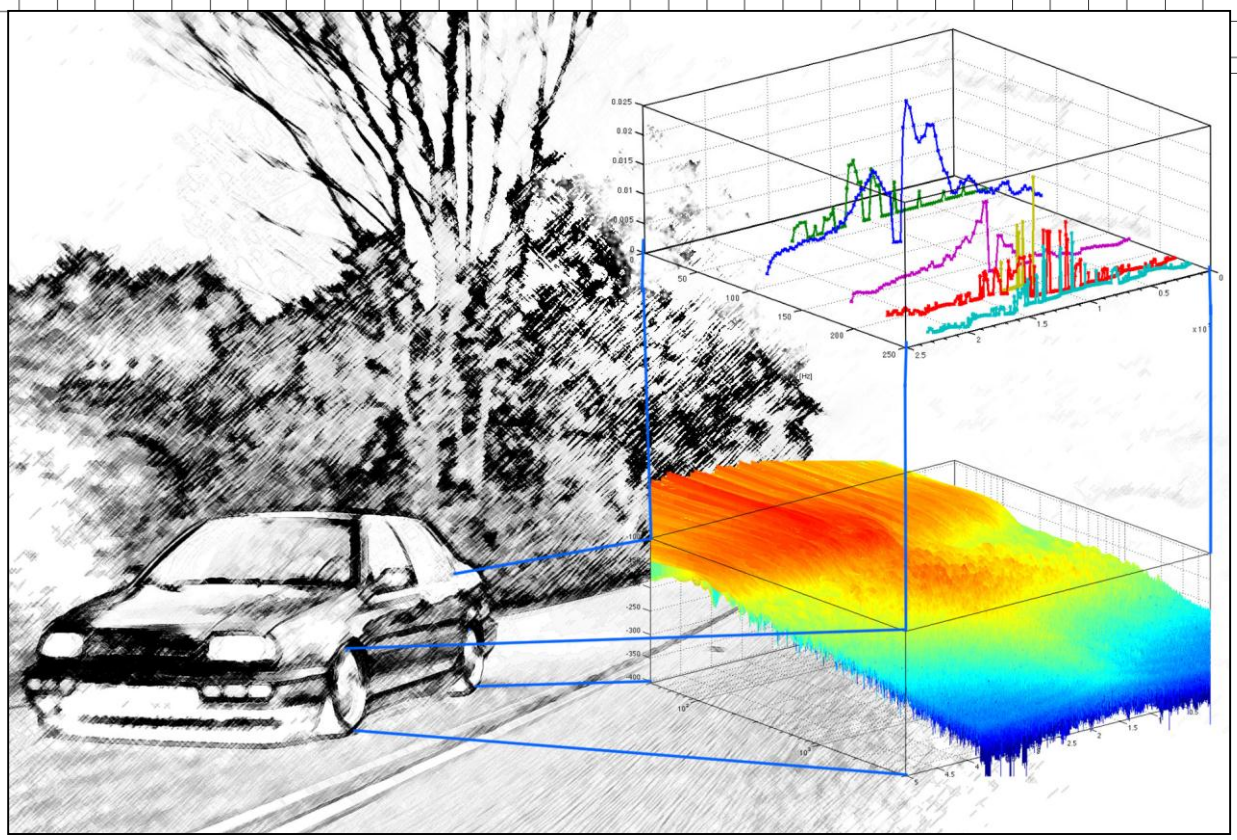
CHALMERS



Auralization of road vehicles using spectral modeling synthesis

Master's Thesis in the Master's programme in Sound and Vibration

CHINMAY PENDHARKAR

Department of Civil and Environmental Engineering
Division of Applied Acoustics
Vibroacoustics Group
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2012
Master's Thesis 2012:154

MASTER'S THESIS 2012:154

Auralization of road vehicles using spectral modeling synthesis

Subtitle

CHINMAY PENDHARKAR

Department of Civil and Environmental Engineering
Division of Applied Acoustics
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2012

Auralization of road vehicles using spectral modeling synthesis
Master's Thesis in the Master's programme in Sound and Vibration

Master's Thesis 2012:154

Department of Civil and Environmental Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-41296 Göteborg
Sweden

Tel. +46-(0)31 772 1000

Reproservice / Department of Civil and Environmental Engineering
Göteborg, Sweden 2012

Cover :
Auralization of Vehicle Sounds by Sayanee Basu, 2012
based on October Outing by Ian Alexander Norman, 2009

Auralization of road vehicles using spectral modeling synthesis
Master's Thesis in the Master's Program in Sound and Vibration
CHINMAY PENDHARKAR
Department of Civil and Environmental Engineering
Division of Applied Acoustics
Chalmers University of Technology

Abstract

The LISTEN Project is collaborative project aiming to build a demonstrator software system for simulation and auralization of the sound environment of road and rail traffic of an urban area. The purpose of the demonstrator is to be able to listen to an urban soundscape at the planning stage, and also allow one to hear how noise prevention measures, such as a noise barrier, will affect the soundscape.

Auralization is used in the LISTEN project to simulate the desired soundscapes. The ability to recreate on-demand soundscapes of given environments allows designers to understand the limitations and practicalities of noise abatement and control.

This thesis looks at using Spectral Modeling Synthesis (SMS) to model vehicles as sound sources to be used in the auralization, along with outdoor sound propagation techniques, to allow simulation of urban soundscapes for the LISTEN project. The SMS technique has its roots in the computer music field, where it was originally used for digital creation of musical sounds using electronic synthesizers. The flexibility of this synthesis technique has allowed it to be used for modeling of various types of sound sources, including traffic and vehicle sounds for study of outdoor sound propagation.

This thesis implements and analyzes the use of SMS techniques for use in modeling road vehicle noise. The Analysis/Synthesis pair of operation defines the core of the SMS system, modeling and auralizing the sound respectively. The perceptual similarity between the original sound and the synthesized sound is also compared in a listening test for analysis of the use of this technique in auralization of road vehicle sounds.

Keywords: Outdoor Noise Propagation, Audio Synthesis, Spectral Modeling Synthesis, Auralization, Sonification, Sound Synthesis, LISTEN Project.

Contents

Acknowledgements

First and foremost I would like to thank my supervisor at Chalmers University of Technology, Associate Professor Jens Forssén for supporting me on this thesis work. I am thankful for the great help, advice and guidance he has given me over the course of the thesis work. His patience and steady support has been critical in helping me learn and understand more in the area of Auralization and outdoor noise propagation.

As this thesis work was a part of the LISTEN project, I would like to thank the members of the LISTEN project for giving me an opportunity to work with them and learn from their experience in this field. Specifically, I would like to thank Peter Lundén at the Interactive Institute for his guidance on synthesis and auralization as well as Mats Nielsson at the Gösta Ekman Laboratory of Stockholm University for his advice on the design of the listening tests.

I would also like to thank the staff at the Division of Applied Acoustics at Chalmers University of Technology for the conducive and supportive environment I was given to focus on the thesis and also the resources that were needed for the thesis work and the listening tests.

Specifically within the Division of Applied Acoustics at Chalmers University of Technology, I would like to thank Penny Bergman for her help with the listening test data analysis, Professor Mendel Kleiner for his feedback and suggestion on the thesis work and Professor Wolfgang Kropp for his inspiration and guidance throughout the thesis work.

Finally, I really appreciate the help of fellow student of the Division of Applied Acoustics at Chalmers University of Technology on various topics through-out the thesis work through the form of arbitrary discussions or presentations. I also am very thankful to all my friends and other students at Chalmers University of Technology who helped with the listening test.

1. Introduction

1.1. Background

Sound surrounds us in all our lives. As a society, we live in a world engulfed by many types of sounds, generated by various aspects of our lives like transportation, commerce and entertainment. These sounds form an environment within our society as well as affect and interact with various segments of our lives.

While some parts of this sound environment in our urban landscape are critical for our daily lives, most parts of this environment are a serious threat to our health, both physical and psychological, and furthermore, impair the possibilities of recreation. In particular, traffic noise pollution is a great and increasing environmental problem.

Traffic noise has been related to sleep disturbance, and also certain cardiovascular conditions [van der Aa 10]. Due to the adverse health effect of traffic noise, the World Health Organization (WHO) has recognized environmental noise, including traffic noise, as a serious threat to public health. A conservative estimate of the health costs of environmental noise are in the range of 40 Billion annually in Europe. WHO claims that more than 40% of the European population is exposed sound levels exceeding the maximum levels published by WHO, which indicates a serious annoyance and health hazard.

Over the past decades, significant amount of research has been done on ways to reduce the environmental noise and its effect, especially within the urban environment. It has been shown that soundscapes of an urban environment can be made more pleasant and healthy if the planners and designers of these environments can be made aware and are empowered. There needs to be a shift in focus from a quantitative perspective of noise, considering only the level of the noise, towards a qualitative perspective where the observer is considered rather than just an acoustical quantity. However, the qualitative aspects of a sound environment are difficult to communicate, particularly to people without special training. Many times groups of people who can have a major influence on the sound environment through their decisions, e.g. politicians, governmental administrators, experts from building industry and citizens, do not have the required training to be able to measure and communicate the quality of sound. The traditional techniques used to document and communicate sound environments are noise maps, which are rather abstract and difficult to understand for the common population.

To aid in the means for communication of the qualitative aspects of a sound environ-

ment, there is a need for a tool, some methods and terminologies which allow effective transfer and understanding of it. There is an urgent need for a better and more understandable representation of the sound environment.

The LISTEN Project (see e.g. [2]) aims to study these issues and create a software system which is able to demonstrate and allow the stake holders to listen and understand the qualitative aspects of specific soundscapes and the corresponding health effects in particular urban environments.

1.1.1. The LISTEN Project

The LISTEN project is a three year research collaboration between the Interactive Institute, Kungliga Tekniska Högskolan (KTH) -Marcus Wallenberg Laboratory, University College of Arts, Crafts and Design (Konstfack), Stockholm University - Gösta Ekman Laboratory and Chalmers University of Technology - Applied Acoustics.

The goal of the project is to build a demonstrator software system for simulation and auralization of the sound environment of an urban area. The purpose of the demonstrator is to show that it is possible to listen to an urban soundscape which is still at the planning stage of development, and also allow one to hear how noise prevention measures, such as a noise barrier, will affect the soundscape.

The project focuses on the simulation of road and rail traffic within urban environments. Various scenarios are considered in this project including various types of road and rail noise, existence and non-existence of noise barriers and positioning of the receiver in backyards as well as inside apartment rooms.

This thesis is a part of the LISTEN project looking at a part of the auralization of sound sources for the urban soundscape.

1.2. Auralization and Audio Synthesis

The term Auralization was first coined by Kleiner [Kleiner 93] to mean the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space. The ability to recreate listening environments has always been one of the aims of acoustics and audio engineering. The goal has been not only to recreate the sensation of the speech or music, but also allow the recreation of the aural impression of the acoustic characteristics of a space, be it outdoors or indoors.

Auralization (also known as Sonification or Virtual Acoustics) is a very important part of the acoustic designers toolbox and part of a solution for a better sound environment. The ability to recreate on demand any arbitrary soundscapes, allows designers to understand the limitations and practicalities of noise abatement and control. Furthermore,

the ability of Auralization techniques to allow the qualitative listening of the soundscape makes it a useful tool.

In the case of the LISTEN project, Auralization is utilized to model the sound sources and simulate the propagation of the sound in the urban landscape for each of the scenarios. Hence, the source modeling and propagation can be considered two of the major areas where Auralization techniques are employed in the LISTEN project.

Modeling of outdoor sound propagation is a matured field of study, with multiple analytical and numerical methods for obtaining the relationship between sound at the source and the receiver. Various environmental effects as well as conditions need to be taken into account to be able to generate these relationships. The propagation studies for LISTEN project were conducted and modeled by Forssén et al [Forssén 09].

The sound source modeling has been an area of interest for many fields of Acoustics. From computer musicians to researchers looking at road and tire noise, the modeling of the source of a sound allows one to have an insight into the inherent properties and mechanism of the source of the sound. Once such models are generated, they can be used to synthesize sounds with various source properties, allowing one to simulate sounds generated in response to various changes to the source. This is known as Audio Synthesis.

The term Audio Synthesis (also referred to as Sound Synthesis) was originally used to describe the digital creation of musical sounds by electronic synthesizers. However, with the advent of modern technology and the adoption of such technologies by the musical community, the term Audio Synthesis has expanded from just looking at musical sounds, to synthesis of sounds of all types.

With the major contribution from the computer music research community, there exist a multitude of techniques of Audio Synthesis, which have been developed over the years each targeted to a specific concept or perspective of looking at sound. Spectral Modeling synthesis [Serra 90], Frequency Modulation Synthesis [Chowning 73], Granular Synthesis [Roads 93], are all different approaches to Audio Synthesis based on various methods of generating and manipulating sounds. Different techniques have been used to model different types of sounds based on the strengths of the technique.

Spectral Modeling Synthesis (SMS) , is a technique which was developed in the early 1990s by Serra, X. at the Center for Computer Research in Music and Acoustics (CCRMA), at Stanford University. This technique was developed for modeling musical instruments and relies on the very tonal sounds made by such instruments. The technique has since been developed further [Serra 03] to allow a wider array of sounds to be modeled, by integrating the abilities to capture various other aspect like noise levels, transients, etc.

This thesis looks at using Spectral Modeling Synthesis to model vehicles as sound sources to be used in Auralization along with outdoor sound propagation techniques to allow simulation of urban soundscapes.

1.3. Purpose and problem definition

1.3.1. LISTEN Demonstrator

The demonstrator created for the LISTEN project is to allow the qualitative listening of urban landscape. The demonstrator has to be able to generate perceptually believable soundscapes, as well as allow the real-time interaction of the user to give a sense of realism [Forssén 09]. These two requirements enforce the direction of the development of the demonstrator.

The demonstrator is implemented in the Pure Data (PD) [Forssén 09] programming environment. The calculations for the propagation model are based on the engineering methods from state-of-art noise mapping prediction, the European Harmonoise [van Maercke 07] and the Scandinavian Nord2000 [Plovsing 06] methods.

The original design of the demonstrator used pre-synthesized vehicle sounds as source for the noise. These were then put through the propagation model to yield the sound at the receiver location. A large number of vehicle sounds were required to have a database of sounds which could be used to generate the various combination of vehicles (light, medium, heavy), and their travel speeds. While this approach allow both the real-time and the perceptual requirement to be met, the resulting database was significantly large (in the case of a preliminary test on just light vehicles, the database size was 5GB). And to meet the real-time requirements, this database needed to be loaded up into the system RAM, which became a limiting factor to the system's real-time performance. Hence, the database approach did not scale when looking at a more complicated source model comprising of various types of vehicles. Furthermore, while looking at alternative approaches to the deployment of the system in the future, including a possible web-based implementation, such a huge RAM requirement seemed limiting.

Thus, a synthesis based implementation was sought for such a system, which would allow a faster and a more flexible solution that requires less memory. A hybrid Spectral-Granular synthesis model was also experimented with prior to this thesis project. However, due to limited success in modeling the sounds of a bus, a more traditional approach of using just the Spectral Modeling technique was chosen for this thesis.

1.3.2. Audio playback vs. synthesis

Audio playback and audio synthesis are the two approaches of modeling the source in Auralization. Audio playback just refers to the technique of recording the source in appropriate conditions (in the case of cars, on a specially designed recording track), and use those recordings as source material for the Auralization. The sound propagation model can then be applied to the source recordings based on the source and receiver positions and surroundings, and the final sound at the receiver can be calculated.

The playback approach is simple and elegant, however it does not scale easily if a

variety of sources is required. In the case of the LISTEN demonstrator, it was important to incorporate various types of vehicles into the source so a realistic noise source could be used. Also, when trying to generate a source for simulating vehicles traveling at a combination of speeds, a playback approach gets a little unwieldy with the large amount of recorded data required to be able to play back such a combination source.

A synthesis approach has its own advantages in Auralization. Synthesis models tend to use much less memory than recordings, hence theres an inherent data compression, which allows the model based approach to be a lot more flexible. Also, synthesis algorithms can designed to be able to produce sounds in real-time allowing live interaction of the listener with the demonstrator and yet not use up much computational effort.

Furthermore, a synthesis approach allows the manipulation of the sounds produced, without requiring multiple pre-calculated sounds. For example, in the case of a vehicle noise model, if the speed of the vehicle can be extracted as a parameter of the model, then a single model could be used to generate sounds for that type of vehicle traveling at various speeds. Other manipulations are also possible, depending on the synthesis model and technique used and the ability to extract various parameters from them.

1.4. Problem definition

The main purpose of this thesis work is to develop and tune a Spectral Modeling Synthesis based system for auralizing sounds of a vehicle passage for use in the LISTEN Demonstrator. Using Spectral Modeling Synthesis, implementing an Analysis and Synthesis technique which can be integrated along with the propagation models in to the LISTEN Demonstrator for a realistic and real-time simulation of the urban soundscape.

1.5. Limitations

The system developed as a part of this Thesis work has a few limitations relating to the assumptions made about environmental conditions, and also the sources being modeled.

Since motivation for this research work was dealing with vehicle sounds, it was assumed the content of the sound being modeled being very much related to sounds produced by vehicles, with a strong tonal component (engine contribution) and a wide-band noise (tire contribution). Other types of noises, especially transient noises, were ignored from consideration in the system.

1.6. Thesis structure

The thesis is structured as follows:

Chapter Two: Contains the Theory the Spectral Modeling synthesis technique and how it can be used to Analyze sounds to create sound models and then Synthesize these models back in to sounds.

Chapter Three: Presents the implementation details of the system. This chapter looks at how the various parts of the SMS technique and the additions to it were implemented in the system, as well as the details of the sounds model.

Chapter Four: Covers at the analysis of the implemented system and also the results of tests done to verify the perceptive realism of the synthesized sounds.

Chapter Five: Discusses the implications of the results and the use of such sounds models in the demonstrator system.

Chapter Five: Finally, all the results and findings are summarized. In addition, research topics for continuation of this work are discussed.

2. Theory

This chapter looks at the theory behind the Spectral Modeling Synthesis (SMS) technique used to analyze the audio and synthesize the sound models, specifically for vehicle sounds. Initially, the idea behind Time Frequency Analysis is discussed, looking at how frequency content of an audio waveform can be analyzed. Thereafter, the theory behind the Synthesis of the sound models is described. Following that, energy analysis techniques, which are useful in SMS, are presented.

Spectral Modeling Synthesis is based on the Analysis/Synthesis concept. The original, time domain, audio waveform is Analyzed, and various types information, also known as model parameters, are extracted from it. This information can then be used to recreate or Synthesize back the time domain waveform.

While the SMS technique does not aim to be able to recreate mathematically equivalent sounds as the original, it does attempt to use many simplifications leveraging on human psycho-acoustical models to ignore or reduce redundant data and parameters. Hence the method aims to recreate sounds perceptually similar to the original, which bodes well with the aims of the LISTEN Project.

2.1. Time-Frequency Analysis

The Fourier Transform is a critical signal processing operation in the field of acoustics. The transform operation allows the analysis of frequency contents of signals, which is the basis of many acoustical method and processes. The Fourier transform however, makes a few assumptions which limit its use in many situations. For example, the signal being analyzed is assumed to be repetitive and stationary. However, many interesting acoustical waveforms are not stationary. For example, the waveforms of musical instruments are rarely stationary; they gradually increase in amplitude at the beginning and decay at the end.

Vehicle sounds, cannot be considered stationary in the long term either, as the sounds produced by the engine change as the various components of the engine rotate and move in operation. Similarly the noise generated by the tire and road interaction, is once again not stationary and changes depending on the position of the vehicle and the properties of road underneath it. To add to all this, the sounds of a passage of a vehicle can have various propagation effects, like Doppler-effect, which would cause a frequency shifting and thus making the sound even more non-stationary.

Hence, a technique needs to be used where a long audio waveform can be analyzed in short time intervals, during which the assumption of stationary-ness can be considered valid. This is the genesis of the Time-Frequency Analysis (TFA). Time-Frequency Analysis techniques are used to study the frequency content of non-stationary waveform, such that both time and frequency domain information can be extracted and considered simultaneously to some extent. There are limitation to the accuracy of the Time-Frequency Analysis techniques which is formalized by the Signal Processing analog of the Heisenberg's Inequality called Gabor limit ([Gabor 46]). The Gabor limit (see Eq. 2.1) implies that product of a signal's bandwidth, and duration can not be less than one.

$$W_B T_D \geq 1 \tag{2.1}$$

where, W_B is a measure of bandwidth (in hertz), and T_B is a measure of time duration (in seconds).

There are many TFA techniques that can be used to analyze non-stationary data. Among the simplest and most straightforward is the Short Time Fourier Transform (STFT).

2.1.1. Short Time Fourier Transform Analysis

The STFT technique looks at short time instants of the audio waveform that can then be assumed as stationary and used to generate frequency information by the Fourier Transform methods. This yields a complex spectrum of that short time signal. The spectrums magnitude and phase defines the frequency information of that short time signal, thus giving some information about the frequency content of original waveform that specific time instant.

When implementing the Fourier Transform, usually, a window function is used to scale the audio waveform being analyzed to simulate the repetitiveness of the signal. This stops the transform function from generating inexistent high frequency components in the spectrum because of the sudden change at the edge of the signal. Similar windowing has to be done when implementing the Fourier Transform part of STFT.

However, since the original waveform is being split into multiple short time instance waveforms, the windowing functions can also cause any information or local fluctuations (higher frequency components) at the edge of the window to be scaled down and hence lost in the analysis. So, to generate the STFT for a long waveform, the sliding window technique is used to ensure no information is lost at the edges of the short time instant waveforms.

Sliding Window Technique

In Sliding Window Technique, the windowing function is applied to a specific part of the signal thus selecting that short time instant, also known as a frame, to be analyzed. Once that analysis is done, the window is slide (moved to a different time index of the signal) to be applied to another section of the signal. The Fourier Transform analysis of such a windowed frame can be considered to give the frequency content of the frame of the signal centred at the centre of the section of the original waveform that was considered by the sliding window.

The number of temporal indices between the two subsequent locations where the window is applied is called hop length, M. The hop length is a measure of the temporal resolution of the analysis. The smaller the hop length, the closer the time instances of subsequent analysis locations. If the hop length is smaller than the length of the window, N, there will be an overlap in the audio signal analyzed in the subsequent Fourier Transforms. In cases where the hop length, M is close to the half of the length of the window N, the audio signal at the edge of one window, will end up near the centre of the window in the subsequent analysis, hence ensuring that all the information is captured in one of the frames of STFT. Figure 2.1 shows an example sliding window over an arbitrary audio waveform. The various overlapping windows indicated are the sections of the audio waveform that are analyzed during each hop.

Eq.2.2 gives a discrete-time representation of a STFT, with $x[n]$ being the audio signal being analyzed, $w[n]$ being the sliding window used to limit the analysis to the time instant under consideration. The exponential term implements the Fourier Transform.

$$STFT\{x[n]\} \equiv X[m,\omega] = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{2.2}$$

This expression gives the frequency content $X[w]$ at each time instant, defined by the index m of the audio signal.

A spectrogram is a 3-dimensional representation of a STFT analysis. The MATLAB *spectrogram* function allows one to visualize the frequency content of the signal at each time interval, based on the parameter supplied.

2.1.2. Overlap and Add Synthesis

The STFT technique is used to analyze the frequency content of a time domain waveform. However, due the use of the Sliding Window Technique, while the frequency information at a certain given time instant is captured, there is no simple way to recompose this information back into the original time domain signal.

The Fourier Transform converts a time domain waveform into a frequency domain spectrum. The Inverse Fourier Transform allows the conversion of a frequency domain

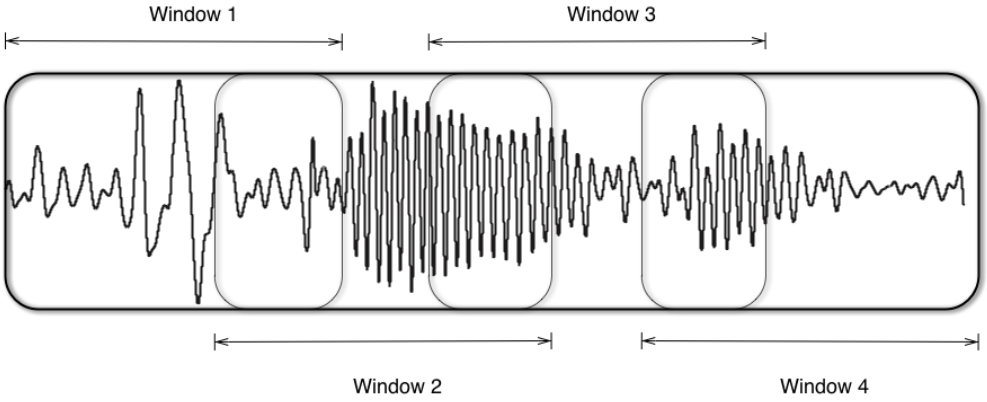


Figure 2.1.: Diagram of a Sliding Window Analysis over an audio waveform.

spectrum into a time domain signal. While these operation are unitary, the overlapping of the analyzed signal caused by the sliding window technique with a hop length less than the window length, does not allow the time domain signal generated by the inverse Fourier Transform to be simply concatenated.

If the two successive STFT analyzed frames undergoes an Inverse Fourier Transform, the overlapped portions of the two successive time domain signals should have the same values, and hence one of them can be ignored allowing concatenation of the rest of the re-transformed time domain signal. However, if some processing and operations are done on the analyzed frequency content, the overlapped portion might not be the same, and hence needs to be considered.

A technique called Overlap and Add is used to recreate the time domain signal. This technique overlaps the successive time domain signal data generated by the Inverse Fourier Transform, scales them by a windowing function and adds the two waveforms. This allows the information from both the successive time domain signals to be captured and yet ensures that the total magnitude of the resultant signal from the addition does not exceed the maximum. This is similar to taking the weighted average of the contribution from each of the time domain signals, and thus captures relevant information.

Constant Overlap and Add (COLA)

While using an Overlap and Add technique for the synthesis of the STFT analyzed audio, one also needs to ensure that the addition process does not affect the signal. The window and overlap ratio need to be constrained to ensure that time domain signal after the, sliding window analysis and overlap-add re-composition, is the same as the original time domain signal.

Considering a sequence of windowed short time instant signals, $x_m(n)$, which are generated from an original signal $x(n)$ using a sliding window $w(n)$ with hop length R, we can enforce the re-composition using overlap and add operation, $x_r(n)$ on that sequence to yield back the original signal $x(n)$.

From Eq.2.3-2.4, this constraint can be met if the sum of the window magnitudes equals to 1. This is the Constant Overlap and Add (COLA) constraint. Various combinations of window and overlap ratios meet this constraint. Table 2.1 gives an example of few such combinations, where R is the hop length and M is the window length.

$$
\begin{aligned}
x_r(n) &= \sum_{m=-\infty}^{\infty} x_m(n) \\
&= \sum_{m=-\infty}^{\infty} x(n)w(n-mR) \\
&= x(n) \sum_{m=-\infty}^{\infty} w(n-mR) \qquad (2.3)
\end{aligned}
$$

Thus, to recompose the original signal after the Overlap and Add,

$$x_r(n) = x_(n)$$

Hence

$$
\begin{aligned}
x(n) = x(n) \sum_{m=-\infty}^{\infty} w(n-mR) \\
1 = \sum w(n-mR) \qquad (2.4)
\end{aligned}
$$

Window Type	Conditions
Rectangular	$R = M; R = \frac{M}{2}$
Bartlett	$R = \frac{M}{2}$
Hamming	$R = \frac{M}{2}; R = \frac{M}{4}$
Hann	$R = \frac{M}{2}$

Table 2.1.: Combinations of window types and overlap ratio which ensure COLA.

Ensuring COLA allows the use of the sliding window technique for STFT analysis as well as the Overlap and Add technique for synthesis without loosing any information.

2.1.3. Spectral Modeling

With COLA, the STFT technique for Analysis and Overlap and Add technique for Synthesis, make up a chain of processes that allow exact synthesis of the original audio. This is the foundation of Spectral Modeling. With the ability to analyze the frequency content of audio waveform and then synthesize it from the analyzed frequency content, the frequency domain information can be used to extract data and parameters.

The Spectral Modeling synthesis technique models the time-varying spectra of an audio waveform as a collection of sinusoids and a time-varying noise component. This

method is based on the Additive synthesis [Risset 85] technique. The Additive synthesis technique assumes that any periodic waveform can be created by the addition of a set of sinusoids at various amplitudes and harmonic frequencies. This simple model however breaks down in the case of many natural sounds, where not all of the critical aspects of the sounds are tonal in nature. A sum of sinusoids is capable of modeling tonal sounds very well, however, with less deterministic sounds, like noisy vehicle sounds, or breath noises in wind instruments, the Additive Synthesis model is not able to capture the energy in the noisy portion of the original waveform, as a noisy signal cannot be described by a sinusoid, but instead is defined by a probabilistic power spectral density.

The SMS technique, separates the audio waveform into; the deterministic component, which can be defined by a narrow band quasi-sinusoidal waveform (with time-varying amplitude and frequency); and the stochastic part, which is modeled by a time varying envelope function defining the amplitude of the probability density of the signal.

The separation of the two types of signal contents allows for easy modeling and data compression of the signal, and also a simpler method to manipulate and process the parameters.

Mathematically, this approach can be modeled as shown in Eq.2.5 degenerating the waveform into a sum of sinusoidal signals and an error signal which is the noisy component. Here $s(t)$ is the original audio signal, A_i is the amplitude of each of the N sinusoid components, ω_i is the frequency of each of the sinusoids, and $e(t)$ is the error signal which is modeled as the stochastic part.

$$s(t) = \sum_{i=1}^{N} (A_i \times \cos(\omega_i \times t)) + e(t) \tag{2.5}$$

The model assumes that all the sinusoids are stable, and the amplitude and frequency only changes gradually, and within a small range, in comparison to the frequency of analysis instances. Such sinusoids usually make up the tonal sounds from musical instruments, and support musical processes of vibrato, and pitch bending, where sinusoidal signal gradually change their frequency within a small range. Such effects are also noticeable in environmental sounds like roughness of engine noises, and the pitch shift caused by the Doppler effect. Hence, this assumption holds well in the case of sounds being modeled for the LISTEN Project.

Figure 2.2 shows a block diagram of the analysis part of the Spectral Modeling Synthesis technique. The analysis part consists of a calculation of a series of spectra based on the sliding window technique and STFT as discussed in section 2.1.1. The spectral information from that operation is used to detect quasi-sinusoidal components in the sound. The spectral content that can not be modeled as sinusoidal components is then captured as a residue and modeled as noise.

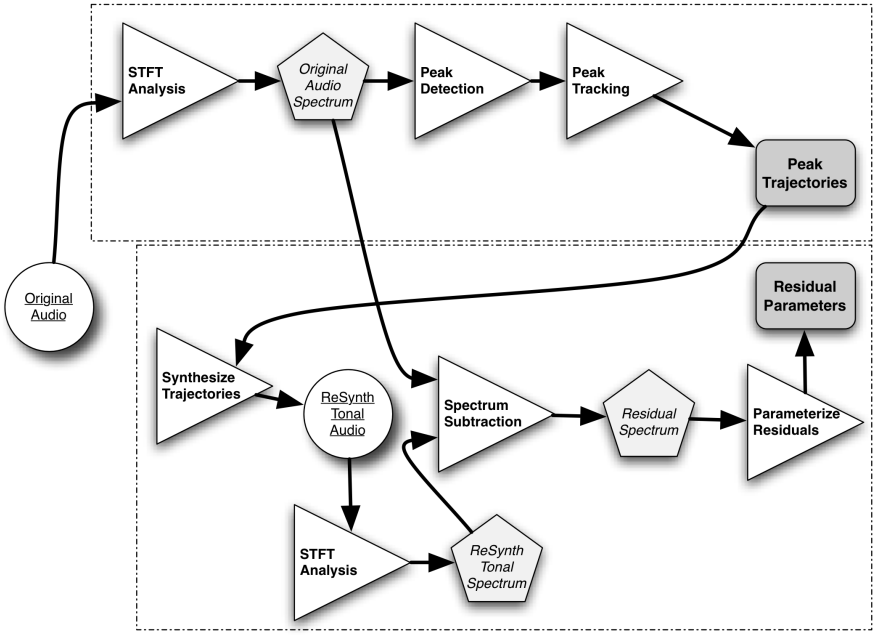


Figure 2.2.: Block diagram of the analysis section of the Spectral Modeling Synthesis Technique.

2.1.4. Peak Detection and Tracking

Peak Detection

A Fourier Transform of an infinitely long sinusoid of a certain frequency yields a Dirac Delta function. In the digital domain, an impulse at the frequency of the sinusoid is seen in the Discrete Fourier Transform of such a sinusoidal signal. Performing STFT analysis of a long waveform, assuming that it is made of gradually changing sinusoidal components, implies that within each frame analyzed by the STFT process the amplitudes and frequencies of each sinusoid are constant, and can be detected from the peaks in spectral waveform. Hence, peak detection is critical in capturing the sinusoidal components in the SMS method.

A peak is defined as a local maximum in the magnitude of the spectrum. Thus, a simple maximum detection algorithm can be used to detect the peaks. However, the nature of the analysis technique makes the peak detection more complicated than just a search for change of gradient of the magnitude spectrum from positive to negative.

Firstly, since the Discrete Fourier Transform (DFT) is used in STFT, the frequency information in the spectrum is also discretized, it is harder to guess the exact maxima of the peak. While a change of gradient, as described in Eq.2.6 does indicate a peak in the vicinity of the discrete points, the data needs to be interpolated to find the exact location (which could be in-between two of the discrete frequency points generated by the DFT) of the peak. A simple way to improve the accuracy of this is to add zero padding when doing the DFT. This causes the difference between discrete frequency indices to reduce, thus giving a better resolution of magnitude, allowing for more accurate peak detection.

$$a_n = x[n]; f_n = f[n]; \forall n \in \{\dot{x}[n-1] > 0 \text{ and } \dot{x}[n] \leq 0\} \tag{2.6}$$

where a_n is the amplitude of the peak and f_n is the frequency of the peak, defined at all points in the spectrum where the gradient changes.

However, even with zero padding, its rare to have the one of the discrete frequency indices to fall exactly at the peak frequency, hence an interpolation still needs to be done to guess the exact maxima of the peak. The shape of the peak is dependent on the analysis window type used in the sliding window technique. Eq.2.7 shows interaction of the window function with the sinusoid, giving the final spectrum a shape of the Fourier Transform of the window itself.

$$FFT(x[n] \times w[n-M]) = W(k - w_0) \tag{2.7}$$

Knowing the shape of the window spectrum, we can interpolate the DFT data to fit that function. An analytical expression of the shape of the window spectrum is needed for such interpolation, which might not be possible for some window types.

However, a good trade off can be achieved by doing a simple quadratic interpolation in the decibels (dB) scale. While this yields an accurate representation of the spectrum function in case of the Gaussian window [Serra 90], a quadratic function can also be a good second degree polynomial estimation of many of the peak spectrum functions, and hence is good enough to estimate the maxima in many cases.

Secondly, the perceptual importance of peaks also needs to be taken into account. Auditory Masking effect makes tones close in frequency to other tones or noise less audible. Thus, in the peak detection stage, it is not necessary to detect all peaks, but instead consider the peaks in the context of the surrounding peaks, picking only those that would be distinguishable to a human hearing.

Finally, in the case of two sinusoid at frequencies very close to each other, the peaks can overlap and make it harder to detect the separate peaks and hence can make the interpolation for the estimation of the spectrum maxima harder as well.

Differential Peak Amplitude Calculation

A more accurate method for detection of peaks was introduced by Desainte and Marchand [Desainte-Catherine 00]. This method uses a DFT of the 1st derivative of the windowed audio signal to calculate a more accurate amplitude and frequency of the peak. Desainte and Marchand show that the ratio of the DFT of the windowed audio signal and the DFT of the 1st derivative of the windowed audio signal are related by a factor of the frequency of the peak as shown in Eq.2.8. This can also be seen from the purely complex Fourier transform, also know as the '$j\omega$ transforms' that is commonly used in physical Acoustics as the $j\omega$ factor relating the Fourier representation of the time domain signal and its first derivative.

$$f_p = \frac{1}{\pi} \frac{DFT^1(f_p)}{DFT(f_p)}. \tag{2.8}$$

Using this method a more accurate estimate of the peak frequency can be detected, without the need to interpolate over the window spectrum function. This method also allows the calculation of accurate amplitude with the knowledge of a continuous spectrum W of the analysis window used for the DFT. Eq.2.9 can be used for that calculation.

$$a_p = \frac{a_p^0}{W|f_p - f_p^0|} \tag{2.9}$$

where, a_p^0 is an estimate peak amplitude, f_p^0 is the frequency of the estimated peak amplitude, f_p is the accurate frequency of the peak, based on Eq.2.8. The required estimates can be seeded from the maximum detection method described in Section 2.1.4.

Another benefit of this technique is that it removes the effect of the analysis window on the amplitude of the peak. Whereas in a normal gradient based peak detection technique

effect of the analysis window on the amplitude of the peak has to be done separately as explained in Section 3.2.

Peak Tracking

While peak detection gives a set of possible sinusoids corresponding to each peak in each frame, the individual peaks need to be analyzed to find patterns of series of peaks which may form a quasi-sinusoidal component. This peak tracking or peak continuation as defined by Serra [Serra 90], tries to find gradually changing tonal components by looking at peaks in each individual time instant and finding peaks that form a good continuation of the sinusoids in the next time instant.

The technique described by Serra uses the idea of a frequency guide, which is advanced every time instant depending on the peaks detected in the frame from that specific instant, forming a trajectory for the sinusoid. The guide helps to smoothly influence the frequency and the amplitude of the sinusoid to generate a gradually changing sinusoid.

The gradual start and end of the guides is enforced in this technique, along with the ability of these guides to remain dormant. A dormant guide may not track any peaks for certain time instants, but may continue after some time instances. This is provisioned to allow scenarios when noise levels go above the amplitude of the sinusoid and it cant be detected. This is typical in sounds with high noise levels, like outdoor sounds. In such scenarios, the guide can pause and start tracking the sinusoid when an appropriate peak is detected in the future time instants.

In the entire process of peak tracking and continuation, the phase information is completely disregarded. When considering sinusoids, only the starting phase of the sinusoids is important as rest of the samples of the sinusoid have a phase that changes regularly with time based on the frequency of the sinusoid. Hence all the phase information can be conveniently ignored in the deterministic analysis and reconstructed during the synthesis by just knowing the initial phase of each of the sinusoid.

In Serra's design, the initial phase of each sinusoid was assumed to be 0. The main motivation for that is the very small number of guides and sinusoids that are needed to be tracked in the musical sounds, and so interaction between tones at slightly different phases would not affect the synthesized sound as much perceptually. This assumption may not hold for other types of sounds.

The peak continuation technique allows a series of peaks to be curated which can define one of the tonal components of the sounds in the SMS model. The amplitude and the frequency of each of these trajectories form the set of parameters that define the deterministic part of SMS model. These trajectories can be used to generate individual sinusoids, which can then be added to generate the final tonal component of the sound. While this modeled tonal component might not be mathematically equal to the actual tonal component in the original sound due to inaccuracies in peak detection and

perception based simplifications, the aim of the model is to be able to generate sound which are perceptually similar, which this sinusoidal modeling accomplishes.

2.1.5. Residuals and Parametrization

Residual Calculation

Peak detection and peak tracking yields the information about the tonal components of the sound. In the SMS technique a part of the sound is assumed to be of tonal and noisy (quasi-stochastic). Having modeled the tonal component, the noisy component can be computed by just removing the tonal components from the original sound. Hence, Serra refers to the noisy component as the residual. This step needs a little elaboration.

The removal of the modeled tonal component can either be done in time domain, waveform subtraction, or in frequency domain, spectral subtraction. Both methods have to ensure that no information is lost during the subtraction process.

The time domain method only works if the phases of the tonal components are known with accuracy. However, in the peak tracking and continuation method, the phase of the tonal component is lost as only the magnitude of the peaks in considered. And since the modeled tonal components are not analytically similar to the tonal component of the original sound, a time domain subtraction would not work in this case.

The time domain method, can be described by Eq.2.10,

$$r[n] = x[n] - x_r[n] \tag{2.10}$$

where, $x_r[n]$ is the synthesized deterministic signal.

The frequency domain method also faces the same issue with the lack of the phase data of the tonal spectrum. Very little information of the tonal spectrum is known, since only the amplitude and frequency of the maxima of the peaks of the tonal spectrum are captured during peak tracking. The rest of the spectral amplitudes have to be generated, and so does phase information of the entire tonal spectrum.

The frequency domain method, can be described by Eq. 2.11,

$$R_i[n] = \begin{cases} X_i[n] - X_{i,r}[n] & \forall \text{ n where } X_i[n] > X_{i,r}[n] \\ 0 & \forall \text{ n where } X_i[n] \leq X_{i,r}[n] \end{cases} \tag{2.11}$$

where, X_{ir} is the i-th frame of the spectrum of the synthesized deterministic signal.

The relationship between the peak maxima and amplitude of the spectrum around the peak depends on the windowing function. But along with that, the interaction of the various peak functions make generation of the peak amplitudes in the frequency domain based on the peak maxima values a complicated task. Serra instead recommends synthesizing the trajectories in time domain (as described in Section 2.2.1) and then reanalyzing them using the STFT method to yield a complete tonal spectrum. This

not only models the interaction of the peak amplitudes but also ensures that any other processing done during the Synthesis process is captured and used in calculating of the residual component. This way the magnitude spectra of the tonal and original signal can be subtracted at each time instant to yield the residual magnitude spectra. While this technique consumes more steps of calculations and is repetitive, the benefits of frequency domain subtractions make it a worthwhile choice.

The lack of initial phase information of the tonal component is still an issue in spectral subtraction method. However, since the residual component is assumed to be noise with a quasi-stochastic power spectral density, the phase information is redundant in the perceptual representation of the sound. Thus, the phase information can be disregarded and generated randomly during the synthesis.

The residual calculation yields a magnitude spectrum of the residual noise for each frame of the sound. Half of this magnitude spectrum is redundant, since the the spectrum of a real signal is always mirrored. Thus, half of the spectrum can be disregarded while the other half can be further parameterized using some noise perception models.

Residual Parametrization

The human perception of noise has been studied very well. Perception of noise in the human ear is not based on the spectral peaks like the perception of tones or even related to information in individual spectral frequencies [Zwicker 90]. Instead the human ear perceives noise based on the energy levels within a energy band [Goodwin 96]. Thus, instead of modeling the residual magnitude spectrum using individual frequency points, it can be modeled using simpler and more compressed expressions based on energy bands. Two approaches for doing this can be considered.

Serra recommends using a envelope modeling approach. Since the individual frequency information in the magnitude spectrum is redundant, the shape of the envelope can be detected and saved in the form of a simple line-segment approximation or more complicated Linear Predictive Coding based approximation [Markel 96]. The two methods have their own strength and limitations with flexibility and total captured data needed.

Another approach presented by Goodwin uses the Psycho-acoustic concept of critical bands. The human ear processes information based on overlapping frequency bands. It has been shown that within these critical bands, the actual magnitude of the spectrum at the various frequencies is not as important as the energy content within that band. Hence each band can be parameterized using the value of energy content in the band, which can be calculated as shown in Eq.2.12,

$$E(b) = \frac{1}{M} \sum_{k \in band b} |X_i[k]|^2 \tag{2.12}$$

M is the FFT Size, $X[k]$ is the kth component of the FFT spectrum and b is the index

of the critical band being analyzed.

Either of the techniques allow the parametrization of the residual spectrum, for data compression and ease of manipulation and synthesis.

2.2. Synthesis

With both the sinusoidal and residual components parameterized, the SMS based sound model can be synthesized to generate a perceptually similar sound to the original sound. The two components of the model can be synthesized separately and then summed to create the final sound.

2.2.1. Tonal Synthesis

The Tonal Synthesis takes the set of trajectories, each being a list of amplitude and frequency pairs, that define gradually changing quasi-sinusoids. The synthesis technique generates a sinusoid for each of the trajectories and finally uses additive synthesis to sum them together to generate the time domain tonal component of the sound.

However since the sinusoidal amplitudes are only defined at the centre of the each frame, the amplitude needs to be interpolated between the two points, over a hop length, to simulate a smoothly changing sinusoid, and also to avoid clicks and other artefacts at boundaries.

The phase of the sinusoid is also required to transition smoothly between each of the frames to generate quasi-sinusoids, however since the original phase information is not captured, the instantaneous phase has to be guessed and generated based on an initial phase and instantaneous frequency of the sinusoid.

$$\hat{A}(m) = A^{l-1} + \frac{A^l - A^{l-1}}{H} m \tag{2.13}$$

$$\hat{\omega}(m) = \omega^{l-1} + \frac{\omega^l - \omega^{l-1}}{H} m \tag{2.14}$$

$$\hat{\theta}(m) = \theta(l-1) + \hat{\omega}(m)m \tag{2.15}$$

Eq.2.13-2.15 define the interpolation of the amplitude and phase information of a single trajectory. The instantaneous phase at each time instant is just the integral of the instantaneous frequency, which in-turn is just the linear interpolation of the frequencies of the peaks defining each frame. The amplitude at each time instant is just the linear interpolation of the amplitude of the peaks.

$$y_d(m) = \sum_{r=1}^{N} \hat{A}_r(m) \cos[\hat{\theta}_r(m)] \tag{2.16}$$

Eq.2.16 defines the final synthesis equation for a single hop length, where all the trajectories defined over that hop are summed up to generate the sound for that hop. Repeating this for each hop in the sound and concatenating them generates the tonal component of the sound.

2.2.2. Residual Synthesis

The residual component of the synthesis is modeled as parameterized magnitude envelope of the residual spectrum. To synthesize the residual component first, the parameters need to be interpolated into the complete magnitude spectrum.

If the line segment envelope method is used, then the line segments need to be interpolated to have the magnitudes at each frequency index. If the critical bands method is being used, then the square root of energy density of each band can be equally distributed among each of the frequency indices in that band, giving a step-wise estimate of the spectrum. Energy density can be defined as the energy per index in the band.

To generate a frame of time domain signal from each magnitude spectrum, the Inverse Discrete Fourier Transform (IDFT) operation can be used. For a complete and real IDFT of the spectra, phase information needs to be added to the magnitude spectra. Based on the assumption of the residual signal being pseudo-stochastic noise defined by its power spectral density, the phase information is redundant. Hence a random phase can be assigned to the magnitude spectra, and the IDFT of that will give a perceivably similar noise signal to the original residual.

$$Y[k] = \hat{E}[k]e^{j\Theta[k]} \tag{2.17}$$

The interpolated magnitude spectrum can then be multiplied by a random phase spectra to form the complex residual spectra as in Eq 2.17. Where, $\hat{E}$ represents the interpolated spectral amplitude, and Θ is the a vector of random numbers. A uniform random phase which covers the entire range from $-\pi$ to π is sufficient to generate the noise. The IDFT of the complex residual spectra yields the time domain residual noise signal as of Eq 2.18.

$$y_r(k) = \frac{1}{N}\sum_{k=-N/2}^{k=N/2-1} Y[k]e^{j\omega_k m}, m = 0, 1, 2, ...N-1 \tag{2.18}$$

Since the residual information is captured and parameterized for each frame, the synthesis process generates the time domain residual signal for each frame. Based on the duality of the sliding window technique for analysis, which was used to generate the residual spectra in the first place, and the overlap and add technique as discussed in Section 2.1.2, the time domain signals for each frame can be overlapped and added to generate the final synthesized residual sound.

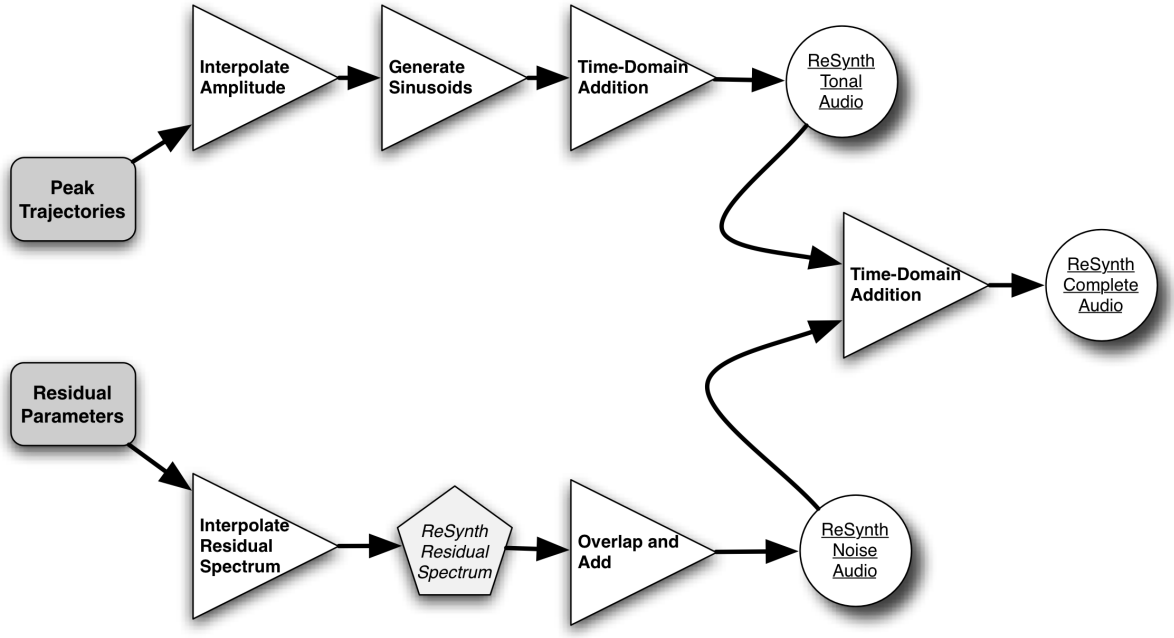


Figure 2.3.: Block diagram of the synthesis section of the Spectral Modeling Synthesis Technique.

Since, the residual being re-synthesized as been assumed as noise, the window function used in the overlap and add process must be one which preserves the perceivable characteristic of the noise, which is power. Serra recommends using a simple Hanning window, but in that case the power gain/loss of the noise as a result of the window needs to be compensated for by scaling the window. This is generally done using a factor known as the incoherent power gain which is defined as:

$$G_{incoherent} = \frac{1}{N} \sum_{i=1}^{N} w[i]^2 \tag{2.19}$$

where $w[i]$ is the window function and N is the window length.

With both the tonal and residual sounds being produced, they again can be summed up together to regenerate the final synthesized sound. Figure 2.3 shows the block diagram for the synthesis section of the Spectral Modeling Synthesis technique.

2.3. Energy Analysis

Another important aspect of the theory behind the Analysis/Synthesis process implements for this Thesis is the Energy Analysis. The SMS technique relies of splitting the given sound into two parts and then recombining them during synthesis. Thus, ensuring that the energy (or power) of the signal is being modeled accurately and tracking the flow of the energy between the various components is critical for such an Analysis/Synthesis process to work effectively.

2.3.1. Energy in Transform Domain

Since the SMS technique uses transitions many a times between times and frequency domains using Fourier Transforms, it is important to understand the calculation of energy in both domains. Parsevals theorem, Eq 2.20, defines the relationship between the energy of time and frequency domain discrete signals.

$$E = \sum_{n=0}^{n=N-1} |x[n]|^2 = \frac{1}{N} \sum_{k=0}^{k=N-1} |X[k]|^2 \tag{2.20}$$

It is important to note the normalization factor of 1/N on frequency domain summation. This normalization factor is also important when dealing with simple Fourier Transforms. A normalized Fourier transform of a sinusoid yields a magnitude spectrum which has an amplitude equal to the amplitude of the sinusoid. This is critical for peak detection in tonal synthesis where the amplitude of the peak is used as the amplitude of the sinusoid.

2.3.2. Critical Bands and Equivalent Energy

Another important energy consideration is during the residual synthesis part of the process. In the method defined by Goodwin, a term of "energy per band" is used to parameterize the residual magnitude spectra. Eq 2.12 is can used to calculate the amount of energy in each of the critical bands. When interpolating, the square root of energy density can be assigned as the magnitude to each of the frequency indices within that band. Taking the square root generates a value in the magnitude domain, while taking the energy density (energy per index in the band) spread the energy equally to each of the frequency index, generating a flat frequency response over that band. Since the perception of the timbre of a noisy sound is based on the total energy of the band [Goodwin 96], spreading the energy equally over all frequencies in the band will not change total energy in the band and thus the timbre of the noise.

The extension of this study of energy during residual synthesis is related to noise distribution. Since the noise being modeled by the residual is assumed be described by an amplitude probability distribution, the nature of the distribution changes the

actual spread of the noise amplitude being generated. When a uniform random phase is applied to the residual spectrum, and then that spectrum is transformed back into the time domain using IDFT, the time domain distribution of the noise generated becomes Gaussian. Thus, regardless of the time domain distribution of the original residual noise, the synthesized noise is Gaussian in distribution. This is visible in spectrograms and other plots where amplitude of the noise signal is plotted.

However, psycoacoustically, since only the energy in the critical bands is perceived in the human ear to understand the timbre, the actual distribution of the noise has no effect on the perception of the noise [Sethares 07]. Hence, both the Gaussian distributed and the Uniform distributed noise sound the same regardless of the difference in the amplitude of the waveform.

3. Implementation

This chapter of the Thesis looks at the implementation of the SMS algorithm for the modeling of vehicle noise for the LISTEN Project. Based on the theory discussed in Section 2, this section looks at the details of the implementation, the changes and additions done the technique to improve its performance. Fine tuning, especially of the analysis part of SMS, was required to effectively and efficiently model the sound.

The implementation of the SMS algorithm, especially the analysis part, uses a number of constants which define the way the data is extracted from the recordings. These constants, or analysis parameters, have to be tuned for individual types of sounds depending on their interaction with the analysis.

3.1. Source Data

The frequency contents of the source data, and other characteristics help to set up some analysis parameters required for the analysis of the sounds. Since the aim of this thesis project was to find good source models for vehicle sounds, a series was vehicle passage recording were used as source data to generate the models. The vehicles passing along a road were measured at a fixed point beside the road, hence recordings were said to be of a passage of the vehicle.

Table 3.1 shows the vehicles used for the development and analysis of the SMS algorithm during this thesis project. The recordings were done largely in accordance with the measurement procedure of ISO 362-1:2007 [ISO 362-1 07].

The four type of vehicle at various speeds provide a variety sounds that can be used for source modeling. The Opel Astra was mainly used as for the comparison of results with another source modeling technique used previously. Having a range of source models

Type	Brand	Model
Car	Opel	Astra
Car	Volvo	V70
Truck	Iveco	Daily (Medium Heavy)
Bus	King Long	XMQ6127C

Table 3.1.: Types of vehicles used for source data.

allows the Auralization of a variety of specific scenarios, for example passage of a bus at regular intervals, or vary the rate and types of vehicles passing by simulating the amount of traffic throughout a day.

The source data was captured into PCM *.wav* files, with a $44.1kHz$ sampling rate and $24bit$ resolution. The files were mostly unprocessed, except when certain occasions when unusual environmental conditions caused sounds not typical of a vehicle passage. For example a small stone on the road, which the vehicle rolled over, or birds chirping in the background could cause a unusual sound to be generated. Since the technique made a few assumptions on the source of sounds, based on the generation mechanism, it was not able to model such unusual sounds accurately. A few recording with such sounds were processed to filter out the unusual sounds.

3.2. Peak Tracking

Peak tracking is the most critical part of the Analysis stage. Peaks need to be accurately detected to model the tonal components. While theory discussed in Section 2.1.4 provides a few methods of peak detection, many improvements can be done to the techniques to make them more effective and efficient.

The combination of a Discrete Fourier Transform (DFT), which is done during the Short Term Fourier Transform (STFT) and the source data, makes the spectrum very peaky. Hence, the detection algorithm finds multitudes of peaks. Some of these peaks in the spectrum are just noise which has a little more energy in a specific frequency component in that specific frame. While the DFT captures the frequency data in that specific frame accurately; considering over a few frames, the energy indicated by that specific peak could be categorized as noise. Thus, it is important to be able to differentiate between important peaks, which indicate presence of tones and less important peaks, which are just a part of the noise spectrum.

As discussed in Section 2.1.4, just the amplitude does not always indicate a the importance of a peak. A variety of other factors need to be considered including neighboring peaks and noise levels.

Noise Threshold

Quasi-stochastic noise is always modeled as having an average sound pressure level, often measured in dB with respect to maximum signal level allowed in the digitized signal. Depending on the timbre of the noise, this level changes over frequency. These noise levels are often considered as boundaries or thresholds. Any signal having amplitude below this threshold cannot be recovered separately from the noise. The auditory masking effect in human audio perception also makes such signals non-perceivable.

In the case of SMS Analysis, there are a couple of noise levels that can be considered

Noise	Treshold
Background (Ambient)	-70dB
Noise Component	-55dB

Table 3.2.: Noise Thresholds used in the SMS Implementation.

and used at the peak tracking stage. Firstly, knowing an ambient and measurement noise level is effective in being able to differentiate between sounds (tonal or residual) which are a part of the recording and sounds which are external noise. A simple peak detection threshold that is set above this external noise level helps to avoid the detection of peaks caused by the ambient sounds. Such a threshold needs to be measured using the setup used for recording the real signal but with the noise source turned off. For the source material used in this thesis project, a recording made with the same equipment but without any vehicle passing by would yield such a threshold.

Furthermore, based on the assumption in the SMS technique of stochastic and tonal components of sounds, if the average level of stochastic noise of the source can be detected or calculated, that too can be used as a lower limit or threshold on the peak amplitude for detected peaks. Depending on source data measurement methods, sometimes it is possible to have data of vehicle passage with the engine turned off, which can be useful in estimating this noise threshold level. Since the engine noise is very tonal in nature, with that turned off, the noise of the tire-road interaction can be assumed to be quasi-stochastic in nature and hence the significant part of the residual component of SMS. Otherwise, a threshold can be set manually, based on inspection of the sound using a spectrogram looking at the level of the noise floor in comparison to maximum level corresponding to the maximum digital signal level. While, the actual noise floor is likely to be frequency dependent, the thresholds across all frequencies turns out to be a good estimate.

For the sounds used to implement and test the SMS technique, the levels used for noise thresholds are in Table 3.2.

Peak Width Thresholding

Mathematically, the width of a peak can be considered as the frequency bandwidth of the 3 dB drop in amplitude. This value is critical in being able to differentiate between a single, wide, peaks and multiple adjacent peaks. Section 2.1.4 discusses peak shape with respect to the Fourier Transform of the window function used. Other factors affect the width of the peak as well. Interaction between multiple tones close in frequency can cause peaks to overlap, thus complicating the peak pattern.

The psyco-acoustical theory [Zwicker 90] also places importance on peak widths, in

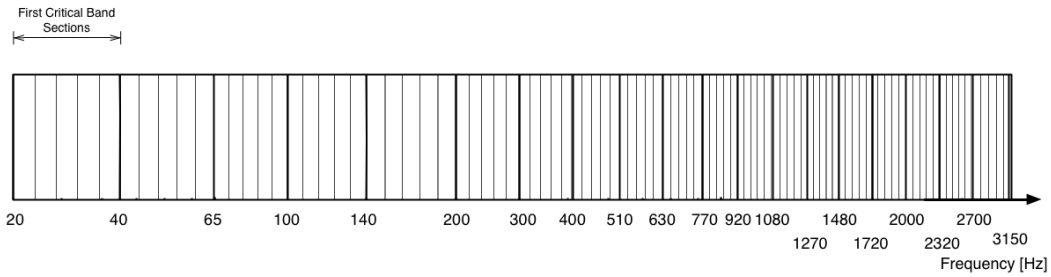


Figure 3.1.: Illustration of Sections in Critical Bands with 5 sections in each band for bands from 20 Hz till 3150 Hz in log scale.

the light of the masking effect. A louder tone may mask off neighboring softer tones as masking effect makes tones of less amplitude within the same critical band non-perceivable. Such tones are redundant to be modeled, as they can't be heard by the listener. Hence, peaks too close to the other peaks can be ignored. Instead, detected peaks should be as far apart as possible from one another to avoid both the miss-detection of wide peaks, and redundant detection of masked tones.

The width of the peak can be calculated based on the window function, and used as a threshold to ensure no other peaks are detected within that range of a previously detected peak. However, the calculated peak widths tend to be a lot smaller than the frequency bandwidth in which tones can mask each other. Since, the threshold for the tones to be masked is not only a frequency bandwidth but also a difference in amplitude between the masker and the tones itself, a accurate pysco-acoustical masking model is complicated to implement and tune for peak detection.

A simpler hybrid solution is proposed in the implementing of the SMS technique for this project. The critical bands are defined by frequency intervals exponentially increasing in width with frequency. While the exponential increase of a full critical band bandwidth models the ears ability to distinguish individual tones very well, it is too large a segment of the frequency to consider for masking. This can be simplified by dividing each critical band into a number of sections. Each section can be allowed to have a single peak, which has the highest amplitude in that section. This ensures that peaks are far enough apart and dont mask each other. Also wide peaks are not incorrectly detected as multiple peaks due to this method. Figure 3.1 illustrates the critical bands and their sections on a logarithmic frequency scale.

The number of sections per critical band is a parameter which has to be manually tuned for the source data being used. A larger number would mean that the allowed

peaks would be closer to each other and could detect some redundant peaks. A smaller number may not be able to track all the tones present in the sound. A number between 5 and 25 was seen to be appropriate in the case of the vehicle sounds. There are ways to automate the detection and tuning of this parameter that have not been explored in this thesis.

Peak Amplitude Calculation

Two methods have been discussed in Section 2.1.4 for the detection and calculation of Peak Amplitude. The quadratic interpolation based method proposed by Serra [Serra 90] and the derivative signal based method proposed by Desainte [Desainte-Catherine 00].

While the division operation between the two magnitude spectra in derivative method cancels the effect of the analysis window on the peak amplitude and frequency; in the quadratic interpolation method, the peak amplitude is still affected by the window function.

A window function is used in the STFT analysis to enforce periodicity on the signal and avoid generation of nonexistent high frequency information. However, a window function reduces the energy in the signal as the signal amplitude is tapered down at the ends. In the transform domain, it can be considered as a convolution with the Fourier Transform of the window function with the spectrum of the signal. This is show in Eq. 3.1,

$$STFT\{x[n]\} \equiv X[m,\omega] = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{3.1}$$

$$= X_m[n] \cdot W[n] \tag{3.2}$$

where, $x[n]$ is the signal being analyzed, $w[n]$ being the sliding window used, $X_m[n]$ is spectrum of the signal and $W[n]$ is the spectrum of the window.

Hence at the peak, the amplitude of the spectrum would be scaled by a factor that is given by the magnitude of the Fourier Transform of the window function at the center of the window. This is known as the coherent gain factor of the window. This factor corrects the amplitude of the peak. This is crucial as the amplitude of a tone in SMS is calculated using the peak amplitude.

The derivative signal based method proposed by Desainte however needs the continuous value of the Fourier Transform of the window function to be able to calculate the accurate peak amplitude. Since the Fourier Transform of the window functions is not always possible to calculate analytically, a high resolution discrete Fourier Transform is used instead. This simulates a continuous valued Fourier Transform, assuming a gradually changing spectrum. This is computationally slow and has to be 'cached' to improve the analysis computational performance.

Both techniques are implemented for the project and can be used interchangeably. A very little difference was seen in the values generated using the two techniques.

Peak Ordering

When the peaks are detected they have to be ordered in terms of their importance for the assignment to various trajectories. The ordering of the peaks can be done based on their amplitude, but as suggested by Serra, X. [Serra 90] importance has to be given to the difference between the peak amplitude and the nearest valley amplitude. Tones with the greatest amplitude with respect to all other tones in that critical band are more perceptible [Zwicker 90], and hence more important in modeling the sound.

A simpler implementation of this is based on ordering of the peaks. Instead of searching for the nearest valley, the gradient of each peak is considered as the metric to order the peaks by. The higher the gradient, the larger the difference between the peak and the nearest valley. The gradient in this case is defined as Eq. 3.3, where $x[n]$ is the signal and n_p is the detected index of the peak. This assumes that the valley is always one index away from the peak, which is true in most cases with noisy signals as in the case of vehicle sounds with the STFT parameters used in this thesis project.

$$G_i = X[n_p] - X[n_p - 1] \tag{3.3}$$

3.2.1. Frequency Guides

Frequency guides are virtual guides that track the peaks for each subsequent frame and form a trajectory of peaks defining amplitude and frequency of quasi-sinusoidal waveform. The generation, decay and advancement of the guides affect the shape of the trajectory and hence the final model. Guides allow gradual change of amplitude and frequency of the sinusoids, thus the parameters and the processes of the guide advancement are critical in being able to track the tonal components well.

Since the original source was a recording of a passage of vehicles, the Doppler effect causes a shift in frequency of the sound at the instance of passage. This shift also has to be detected and tracked accurately in the frequency guides.

Guide State Machine

The virtual frequency guides are given states to allow guides to sleep as discussed in Section 2.1.4. A finite state machine implements the various states of the guides. This allows the state transitions to be defined, and the appropriate actions to be taken when the state transitions happen. Such a process of tracking the guides becomes more useful when further processing is done as the guide finds an appropriate peak while in the sleep

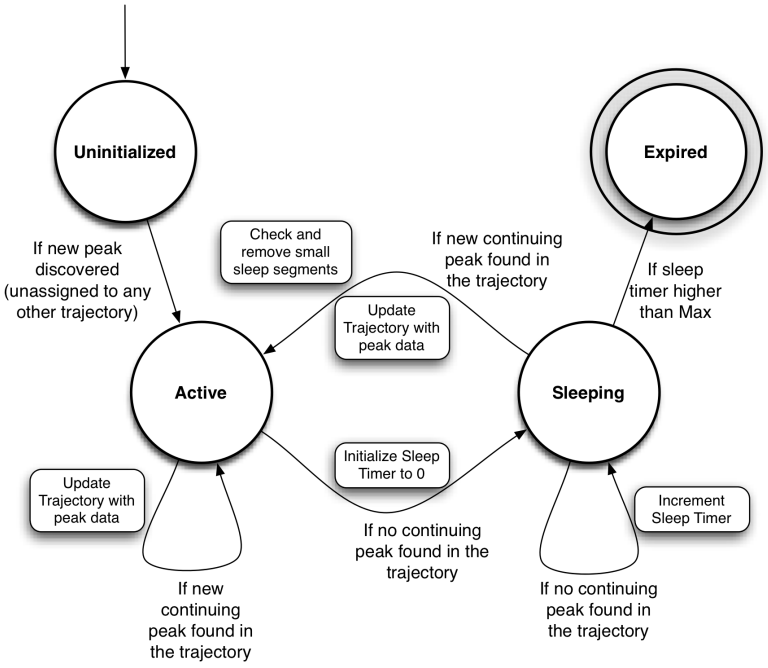


Figure 3.2.: Frequency Guide State Machine Diagram.

or dormant state or as a guide needs to process the tracked tones and smoothen its trajectory.

Figure 3.2.1 shows a state diagram which defines the state machines of the frequency guides. Many state transitions are not allowed and only a few result in specific actions.

3.2.2. Guide Width

Similar to the peak width, the width of a frequency guide is a parameter which needs to be tuned depending on the sound. The width of the guide allows the guide to accept peaks within the range of the width around the center frequency of the guide. A wider guide accepts peaks of frequencies that are far from the current center frequency. This ensures that the guide does not expire quickly as a result of a lack of appropriate peaks. However, a wider guide tends to allow the trajectories to deviate significantly. This might cause confusion with if the sounds have two tonal components that are less than the guide width in frequency apart. In such a case, the frequency guide will tend to jump between the peaks from each of the component, yielding an unnatural sound.

The control of the guide width is critical in being able to detect correct trajectories. Once again the perceptual critical band theory can be used to ensure that the band-with is wide enough to detect differences between perceivably distinct tones, and yet not too

narrow as to not allow the gradually changing tones, like a Doppler shift to be captured within a single guide.

Based on the critical band section method used in peak detection, a similar sections based simplification can be used to define guide widths. Each critical band can be divided into a number of sections; within each only a single frequency guide is allowed to exist. This ensures that the guides are far away from each other and do not deviate too much from the center frequency to track peaks, which may lead them to track peaks from another tone.

Once again, the number of sections per guide is a parameter that needs to be tuned to be able to extract tones accurately. Throughout the implementation of the SMS technique, values between 5 and 20 were used for the number of sections generated from critical bands. A higher number tended to offer better results in vehicle sounds as the tones in the source data did not vary much from the original frequency.

3.2.3. Trajectory Frequency Smoothing

Frequency Guides help to detect and shape a trajectory of peaks. To ensure a gradual change in the frequency of the sinusoid modeled by the peaks, the trajectory frequencies need to be smoothed. This is done as proposed by Serra using a current trajectory bias which is based on a simple low pass filter expression. The progression of the guide is based on a weighted sum of the current trajectory of the guide and the new peak being tracked, thus smoothly shifting towards the new peak.

Eq. 3.4 shows the expression used to calculate the influence of new peaks on the trajectory, where $\hat{f}_i$ is the estimate frequency of the next peak based on the current trajectory, and g_i is the frequency of the next peak being tracked, and f_i is the final frequency assigned to the guide after the smoothing.

$$f_i = \alpha(g_i - \hat{f}_i) + \hat{f}_i \tag{3.4}$$

The constant α is used to decide wether the current trajectory is given more importance or if the new peaks frequency is given more importance when deciding the frequency of the trajectory for the next frame.

This parameter α allows the smoothness of the frequency guide to be adjusted. An α with a value close to 1 favors the frequency of the new peak, while an α with a value of 0 favors the extrapolation of the current trajectory. α is also a parameter which has to be tuned for accurate analysis of the model. Mostly in the implementation of the SMS Technique, an α value of between 0.25 and 0.4 was used. A moderate value of alpha allowed a smooth tracking of peaks and yet still track sudden large changes in frequency like seen during a Doppler shift at the point of passage.

3.2.4. Trajectory Amplitude Interpolation

Along with a smoothly changing frequency of the trajectory of the tone, the amplitude of the tonal component is also assumed to be smoothly changing. Hence, the trajectory amplitudes are interpolated to give a smoothly changing amplitude value for the sinusoids between individual peaks.

Although this interpolation yields a smooth change between subsequent frame, the amplitude of the peaks themselves can vary significantly and might change drastically if no peak is found in the current guide, forcing the guide to go into "Sleep" mode. The negative effect of this is having Frequency Guides that sleep and awake in rapid succession. The Auralization of such trajectories where peaks often have zero amplitude, have artifacts that sound unnatural. This behavior was seen often in signals with amplitudes that are close to the noise threshold.

To reduce these artifacts, a trajectory is analyzed at the end of the peak tracking operation and short sleep periods of Frequency Guides are removed and replaced with guessed values. The definition of short is again another parameter that needs to be tuned. However, the value of this period may be calculated, based on the time interval needed to perceive the decay of a tone, and thus it would depend on the hop length. Eq. 3.5 shows the expression which could be used for this,

$$N_{limit} = \lfloor \frac{T_{min} \times f_s}{M} \rfloor \tag{3.5}$$

where M is the hop length, T_{min} is the minimum time for a sleep period not to be interpolated, and f_s is the sampling frequency.

Care needs to be taken though as interpolating over a larger time period may add energy to the tonal component that might not exist in the original sound causing inaccuracies in the model.

The removal of the sleep period can be implemented by setting the peak amplitude for the frames during the sleep period to values based on simple linear interpolation between amplitudes of the last peak before the Frequency Guide went into sleep state and the first peak after it comes out of sleep state. These allows the tone to change gradually between the two states and not auralize artifacts caused by rapidly changing amplitudes.

For the SMS Implementation a nominal value of 0.5 seconds was considered as the minimum sleep duration, and any sleep period less than that was removed using the interpolation described above. Eq .3.6 gives the formula used for the interpolation,

$$A'_i = \frac{(A_{first} - A_{last})}{(first - last)} \times i + A_i \tag{3.6}$$

where A_i is the original peak amplitudes before interpolation, A_{first} is the amplitude of the first peak after the sleep mode, and A_{last} is the amplitude of the last peak before

sleep mode.

3.2.5. Guide Energy Filtering

Noise thresholding as defined in Section 3.2 helps to reduce the number of redundant peaks. A smaller number of peaks allows a quicker tracking and Auralization of peaks. However that has to be balanced with having very little or no perceptual change the sound. It was observed during testing that sometimes lesser peak trajectories yielded sounds perceptually more similar to the original than if more trajectories were used, as many of the extra peaks tracked and auralized were not from the tonal component but noise peaks which were mistakenly tracked as tones.

Thus, it is useful to remove trajectories that do not contribute much to the final sound from the rest of the analysis. This can be done by a simple energy comparison. A value representing the energy of each peak trajectory can be calculated using the Eq. 3.7,

$$E_j = \sum_i A_{i,j}^2 \times M_j \tag{3.7}$$

where, $A_{i,j}$ is the i-th peak in the j-th Frequency Guide, and M_j is the hop length of the the j-th guide.

While the value may not reflect the exact amount of energy in the sound synthesized by the trajectory, it is the relative value between the trajectories which can be compared to decide which trajectories can be ignored.

A simple percentage of energy comparison can help in the detection of redundant guides. Any trajectories with total energy less than a specific percentage of the combined energy in all the guides, can be ignored. This is another parameter that has to be tuned to ensure that the appropriate number of trajectories should be considered. During the experiments, a value between 0.1% to 0.05% seemed to be appropriate.

3.2.6. Guide limit increments

Even with noise tresholding, there are many peaks that get detected which do not correspond to a specific tonal component. However, since that cannot be verified until no other peaks are found in the subsequent frames, all new peaks get assigned a new frequency guide. This generates a large number of guides, which might not continue beyond a few frames, and hence the total number of guides keeps increasing every frame analyzed. This can quickly get computationally challenging. Hence, a simple limit on the maximum number of guides ensures that only the important peaks can generate new trajectories. And since the peaks are assigned to guides in the order of their amplitudes (or gradient as described in Section 3.2) only the important peaks are assigned before the limit of number of guides is reached.

With a limit on the maximum number of guides, the limit gets reached very quickly within the first few frames. Hence, if an actual tonal component starts midway through the sound, there might not be any frequency guides left to track it. This is not a common effect in musical sounds, where the tonal components are more or less stationary and begin from the start of the sound. But in environmental sounds it is common to have tones that become significant much later in the sound as a result of propagation or other effects.

A proposed solution is to allow additional guides to be added to the tracking algorithm as the further frames are tracked. Eq. 3.8 shows the change of the maximum guide limit can have with more and more frames are added,

$$N_{g,i} = \frac{4 \times N_{g,init} \times c_{inc} \times i}{3 \times N_{hops}} + N_{g,init} \qquad \forall i \in [1, \frac{3 \times N_{hops}}{4}] \tag{3.8}$$

where, $N_{g,init}$ is the limit of the number of frequency guides at the beginning of the peak tracking process, N_{hops} is the total number of hops, c_{inc} is the increment factor, which is the fraction of the $N_{g,init}$ which are added through the tracking process.

This increase of the limit can be stopped nearer to the end as not many new tones start close to the end of the sound. Experiments show that a guide limit increase for the first 75% of all the frames is enough to deal with most late starting tonal components.

Here, the guide limit, or the maximum number of guides allowed is an analysis parameter that needs to be tuned. The tuning of this parameter is a direct trade-off between the quality of model and the amount of computational time required to synthesize the sounds. Similarly, the increment factor is also a parameter that needs to be tuned for the specific sound being analyzed. Complex sounds with new tones starting at various instances through out the sound might need a higher increment factor.

3.3. Noise Synthesis

A number of implementation improvements also need to be done in the Noise Synthesis part of the SMS method. During the Noise synthesis operation, the residual spectrum is first interpolated from the parameters and then assigned random phase before being transformed to the time domain by an Inverse Fourier Transform.

3.3.1. Phase Randomization

Randomization of the phase assigned to the spectrum serves two purposes. Firstly, the frequency domain subtraction method used to calculate the residual, does not generate the phase of the residual spectrum. Hence, there is a need to generate the phase for the spectrum. Furthermore, since the specific phase information is not perceivable in a noise component, it does not need to be stored and can be generated randomly. Serra

suggests using a uniform random variable with values between 0 and 2π or $-\pi$ and π to generate the phase.

However, to generate a real signal from the complex spectrum using the Inverse Discrete Fourier Transform (IDFT), the spectrum has to have a certain format. The complex spectrum of a real signal is always mirrored and conjugated. Thus the random phase can be assigned to the half of the magnitude spectrum that has been interpolated, and the resulting complex spectrum can be mirrored and concatenated to make the complete complex spectrum. Eq. 3.9 - 3.11 explains this operation.

$$\theta_l(n) = \quad 2\pi \times rand(\tfrac{N}{2} - 1) \tag{3.9}$$

$$\theta_r(n) = \quad -\theta_l(\tfrac{N}{2} - 1 - n) \tag{3.10}$$

$$\theta = \quad [0, \theta_l(n), 0, \theta_r(n)] \tag{3.11}$$

Here, *rand* is a random number generator which produces a vector of random numbers between 0 and 1. Scaling it to 2π gives a vector of random numbers between 0 and 2π. The phase vector needs a value of 0 at the 1st index and at the N/2+1st index. The random vector, θ_l is then mirrored, conjugated and concatenated to give the final phase vector, θ. The IDFT can be calculated from the phase vector and the magnitude using Eq. 2.17.

3.3.2. Overlap and Add of Synthesis

The zero padding used in the DFT/IDFT operation for residual synthesis breaks the duality of Sliding Window analysis and Overlap and Add synthesis. Zero padding is used in DFT operations to improve the accuracy of the spectrum by decreasing the difference in subsequent discrete frequencies in the spectrum. Generally, at least twice the length of the analysis window length is recommended as the length of the padded signal.

However, this creates a spectrum of the same length at the padded signal. Without any other process, if this spectrum is transformed back into time domain waveform using IDFT, it will contain the original windowed signal concatenated with the zero padding. However, during the residual synthesis, the phase of the complex spectrum is randomized. The re-transformed time-domain signal does therefore not have any identifiable zeros padding concatenated at the end of the waveform, and is still of the same length as the padded waveform. Figure 3.3 illustrates this for an arbitrary signal. The inverse Fourier transform of the spectrum without any changes yields the original time domain signal with the identifiable zero padding at the end (bottom-left plot). However the same magnitude spectrum when combined with a random phase and inverse Fourier transformed does not give a time domain signal with a clear zero padding (bottom-right plot).

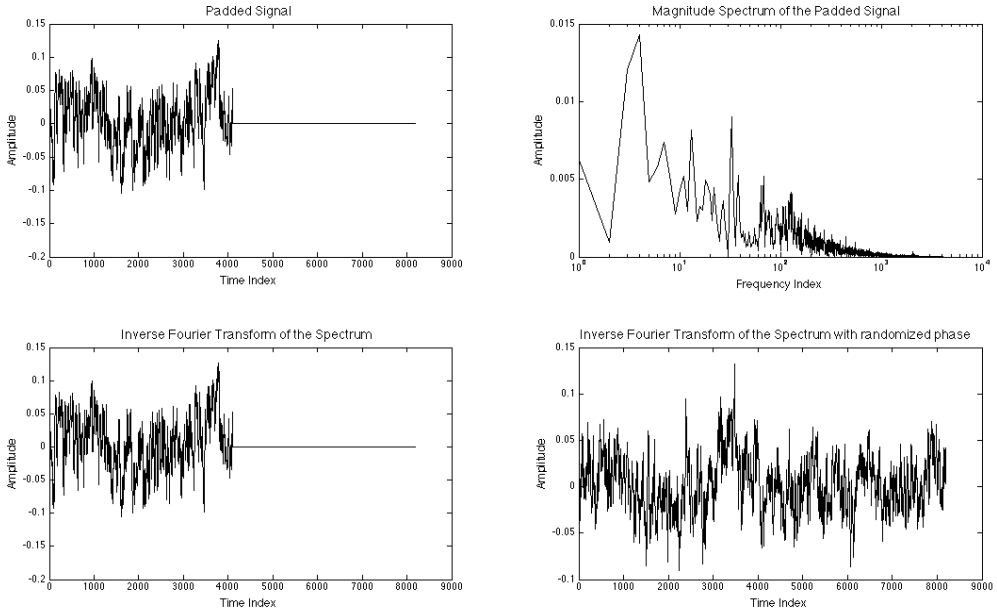


Figure 3.3.: Example plots of (clock wise from top left) an Original Signal, its Magnitude Spectrum, and the Inverse Fourier Transform of the Magnitude Spectrum with and without randomized phase information.

While total energy is still the same, and also at the same frequency (since the spectrum is similar), it is now spread out over the entire signal instead of being concentrated in the initial part of the signal. In other words, since the phase of each of the frequency content in the original signal was discarded and replaced with random phase, the energy in the various frequencies shifted in the time domain.

Since now the signal is longer, it cannot be overlapped and added the same way as the original signal with the zero padding which could be identified and removed before the overlapping. To overcome this, the signal has to be truncated to the length of the analysis window. This truncation of data can cause the loss of information in some frequencies. However, the nature of zero padding does not add any information to the signal and hence all the relevant frequency information can be proven to be captured within the part of the signal which remains after truncation.

Furthermore, since the net energy was spread out over the entire signal, some of it is lost in truncation here, as the truncated part of the signal is not non-zero unlike the case with zero padding. Thus, the signal left after the truncation has to be scaled to have the same amount of energy as original in order to have similar loudness. Here too, the redundancy of the actual noise distribution helps to ensure that the perception of

the type noise is not altered. The scaling can be expressed as defined in Eq. 3.12,

$$x'_r = x_r[1 : N_o] \times \frac{N_z}{N_o} \tag{3.12}$$

where N_o is the original window length, N_z is the window length with the zero-padding, and x_r is the time domain signal of a specific frame after synthesis.

After the truncation and scaling, with a residual synthesis waveform of each frame being the same length as the original waveform, overlap and add synthesis will give an perceptually accurate reconstruction of original signal.

3.3.3. Weighted Overlap and Add

While the COLA constraint (see Section 2.1.2) ensures that a Sliding window based STFT and an Overlap and Add process keeps the original signal constant, it does not take into account the processing of the signal between the two. The process of parameterizing the noise envelope (as defined in Section 2.1.5) affects the noise signal itself. When the residual is originally calculated, the STFT of the original and the re-synthesized tones are used. Since the STFT process windows the signal before they're subtracted, each residual/noise Hop can be considered to be windowed. Eq. 3.13 shows the analog of the frequency domain subtraction in time domain:

$$r_m[n] = x_m[n] - x_{r,m}[n] r_m[n] = x[n]w[n - mR] - x_r[n]w[n - mR] r_m[n] = w[n - mR](x[n] - x_r[n] \tag{3.13}$$

Here, x_m, $x_{r,m}$, r_m are the original, re-synthesized tone and residual signal respectively, of the mth Hop.

The process of parameterization, interpolation as well as synthesis using the random phases (see Section 2.1.5) causes the re-synthesized residual signal to loose it's windowed amplitude. Hence, in order for the overlap and add to work, it needs to be windowed again. Thus the finally re-synthesized noise signal has gone through windowing twice. Hence the COLA constraint cannot be applied to it directly. Instead a weighted version of the COLA constraint, Weighted Overlap and Add (WOLA) as defined by Smith [Smith 12] can be used.

Since the residual signal is windowed twice during the entire SMS process, it has to be constant after the overlap and add process when the square of the window function (same window applied twice) is applied to it. Smith recommends using the same window both the times for ease of finding a WOLA constant combination of window and overlap ratio. Eq. 3.14 defines the new WOLA constraint:

$$r_r[n] = r[n] \sum_{m=-\infty}^{\infty} w[n - mR]^2 \tag{3.14}$$

Thus, to recompose the original signal after the Overlap and Add,

$$r_r[n] = r_[n]$$

Hence

$$r[n] = r[n] \sum_{m=-\infty}^{\infty} w[n - mR]^2$$
$$1 = \sum w[n - mR]^2 \tag{3.15}$$

Here $w[n]$ is the window being applied during the STFT and as well as the overlap and add stage.

Since the combinations of windows and overlap ratios have already been generated (see Table 2.1), the windows which satisfy the WOLA constraint can be easily calculated by taking a square-root of the window functions of each other windows in the table. Taking the Hann window as an example, it's square-root, a Sine Window can be used to satisfy the WOLA constraint along with a overlap ratio of 50% ($R = \frac{M}{2}$).

While, using the Sine window based on the WOLA constraint is very accurate and theoretically correct, it was not used for the implementation of the thesis because of a oversight. Instead a Hamming window with 75% overlap was used. Since this affected the net power of the re-synthesized noise, the effect of was reduced by normalizing the re-synthesized noise with the net power of the Hanning window. While, not mathematically accurate, this method did reduce the effect of double windowing to a great extent and hence it did not create any audible artifacts, especially in term of loudness. Also, further processing done to scale the energy (see Section 3.5) fixed any discrepancies in the energy levels.

3.3.4. Amplitude Smoothing

While testing the synthesis, an artifact was audible in the higher frequencies, which sounded like modulation of the noise amplitude. This artifact made the noise in that specific frequency band sound un-natural. An analysis of the residual parameter in the last few critical bands showed a large fluctuation of the residual parameter in certain critical bands. The residual parameter which controls the amplitude of the noise in that band was fluctuating significantly for every frame it was calculated for. A reason for this could be that the window length being used to analyze the high frequency noise is too small and is being dominated by the local fluctuations of the noise instead of capturing the average values. While changing the window length just for the residual synthesis could be a possible solution, it would require further processing of the residual parameters to make them compatible with the corresponding tonal parameters, making it much complicated and computationally intensive.

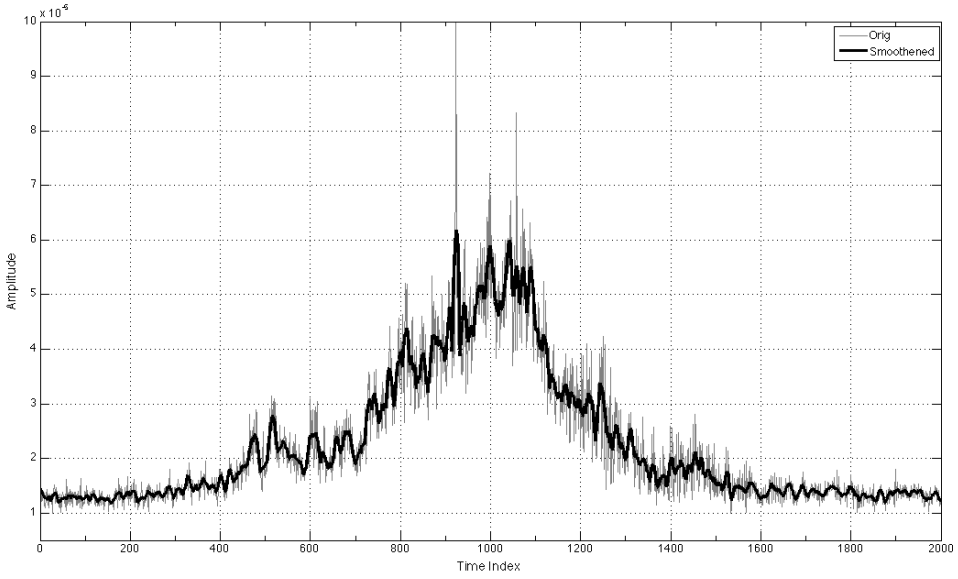


Figure 3.4.: Effect of the smoothing filter on fluctuating Residual Parameter.

Tone Frequency	Samples per cycle	Min Window Length	Max Window Length
40Hz	1102.5	1024	8192
16kHz	2.75	2	16

Table 3.3.: Samples per sinusoidal cycle, and minimum and maximum window length (in the powers of 2) assuming a sampling frequency of 44.1kHz.

A simple solution was to filter the signal formed by the residual parameter for each successive frame through a SavitzkyGolay smoothing filter [Orfanidis 96]. This filter reduces the individual fluctuations of the signal while still keeping overall shape of the noise parameter signal. While a simple low pass filter could have also worked, the SavitzkyGolay smoothing filter is better at preserving features of the signal such as relative maxima, minima and widths, thus reducing the effect on the total energy of the noise. The filter reduces the amplitude modulation artifact and make it less perceivable. Figure 3.3.4 shows the effect of this filter.

3.4. Multi-pass Analysis

In a STFT analysis, the choice of the analysis window length is critical. Too short a window and it might not be able to capture enough samples of a low frequency tone

Window Length	Hop Length	Padded Window Length	Min Frequency	Max Frequency
128	32	512	344.53 Hz	2756.2 Hz
512	128	2048	86.13 Hz	689.06 Hz
2048	512	32768	21.53 Hz	172.27 Hz
32768	2048	131072	5.38 Hz	43.06 Hz

Table 3.4.: Window length, the associated Hop Length, Padded Window Length and corresponding calculated minimum and maximum trackable frequencies assuming a sampling frequency of 44.1kHz.

to be able to detect it. Too long a window and the high frequency component may have gradually changed within that duration thus making that change undetectable to the STFT analysis. Thus the window length always has to be in the range of $\frac{1}{2}$ to 4 multiples of the length of a sinusoid of any frequency needing to be tracked. Since audio signals are being considered, the frequency range of interest for human perception of tonal information is from 20Hz to about 15-16kHz. This range can be truncated for natural sounds that mostly lie within 40Hz to 16kHz. Table 3.3 shows the minimum and maximum window lengths for a 40Hz tone and a 16kHz tone to be detected correctly.

Hence, no single window length can be used to capture the data in the entire frequency range accurately. This can be overcome in a few ways. Firstly, using a non-sinusoidal function as a basis of the transform instead of a Fourier transform can help to be able to capture much larger frequency range with a single window length. Wavelet based transforms [Strang 93] are a good example of this method. However, a wavelet based transform deviates from the fundamental aspect of the SMS technique, making the separation of the tonal component complicated and non-intuitive. This method, although valid, was not implemented for this project; instead a simpler multi-pass method was used to capture the data.

3.4.1. Frequency Bands and Filtering

To allow the capture of tonal components at all frequencies multiple passes for STFT analysis can be done. For each pass the waveform is analyzed with windows of different lengths, thus making sure that all the frequency components get analyzed with a window that can detect them. Since the window length is changed, the length of the zero padding as well as the Hop length has to change to keep the other parameter compatible. Table 3.4 shows the values which might be used for a 3-pass analysis, and the calculated minimum and maximum frequencies that can be tracked based on the $\frac{1}{2} - 4$ wavelength assumption. Keeping to the benefits of using lengths with powers of 2, the various window lengths are also chosen in multiples of 4. This is known as a pyramid STFT

analysis. Figure 3.5 shows an illustration of such a multi-pass STFT analysis.

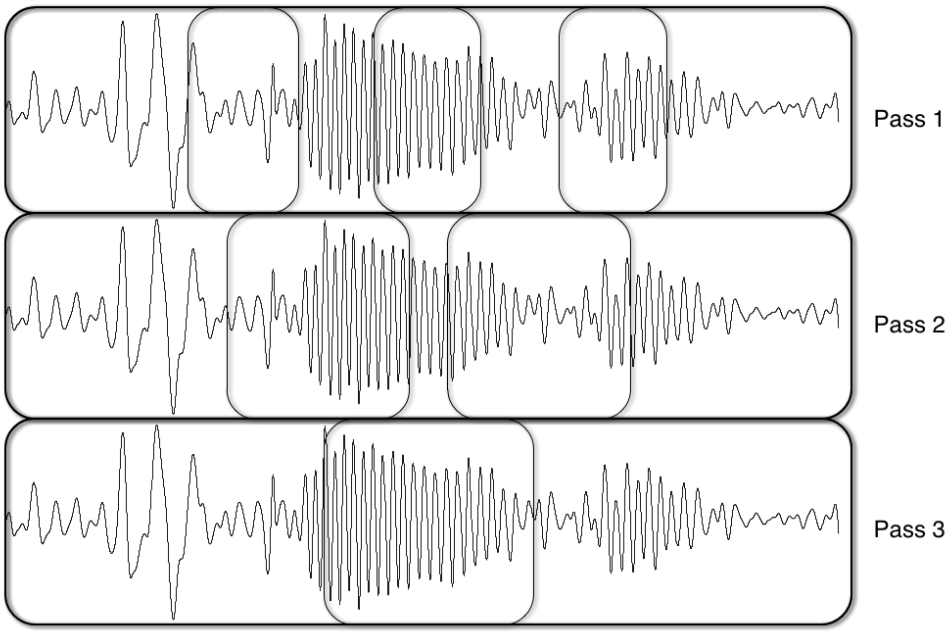


Figure 3.5.: Diagram of a Multi Pass STFT analysis scheme using a sliding windows of various sizes.

As a result of multiple passes, a set of peak trajectories are generated for each pass, with different time index lengths between each. However, to force this separation, frequency based filtering has to be done for the peak detection. This ensures that peaks which might be found during the STFT analysis will not be tracked if they do not lie within the band of the frequency which corresponds to the hop length used for the STFT analysis. Thus, assigning a frequency band for each pass, only peaks within the band are considered in the corresponding pass for peak tracking.

Similar filtering is also done for residual calculation. The peak trajectories of each pass are synthesized separately, and used to calculate the separate residuals for each pass. The spectrum of the original sound has to be filtered using the same filters as used in the peak tracking before subtracting the synthesized peak trajectories. This also helps to ensure that the appropriate window length is used for calculating the residual magnitudes, so that the information relevant only to the corresponding frequencies is captured in that residue. During residual parametrization only the frequencies relevant to the window

length are parametrized using filters based on the frequency-bands corresponding to the window length.

For synthesis, since both the peak trajectories and residual parameters have been frequency filtered for each pass, they can be individually synthesized using the STFT parameter for the corresponding pass and finally just summed to generate the final signal.

After tuning for the source data, the values for hop length and corresponding frequency limits shown in table 3.5 were used.

Window Length	Hop Length	Padded Window Length	Min Frequency	Max Frequency
128	32	512	10000 Hz	22050 Hz
512	128	2048	4000 Hz	10000 Hz
2048	512	32768	180 Hz	4000 Hz
32768	2048	131072	40 Hz	180 Hz

Table 3.5.: Window length, the associated Hop and Padded Window Length and the Filter Band Limits used in the Multi-Pass Analysis.

While this method improves the accuracy of the tonal component detection greatly, it adds a significant computational overhead. And especially with very small window lengths, many more frames are generated causing the time taken for the analysis to increase greatly. Hence a balance has to be achieved between computational effort and accuracy of the modeling. A higher number of passes can yield a more accurate model, but also add more computation overhead. However, since most of the computational effort is used for the analysis stage, the synthesis stage can be optimized to run quicker and possibly run in real time for a demonstrator tool.

3.4.2. Pass based Guide Limit Assignment

An attempt to improve the efficiency of the analysis stage was made considering the time taken for peak tracking. At higher frequencies with a greater number of frames, peak tracking takes a significant amount of time. However, most of the tonal energy is in the lower frequencies and most of the trajectories in the short window length pass have too little energy to be considered in the synthesis.

Thus to avoid redundant analysis for tones in the higher frequency ranges, a variable number of guide limit was assigned to each pass, which segregate based on frequency. The pass with large window length, where most of the tonal energy was located was assigned greater number of frequency guides to begin with, while the pass targeting the higher frequencies was assigned fewer guides initially. This reduced the time as well as memory footprint of analysis significantly.

These variable numbers of guide limits is again a parameter that needs to be tuned

for the various source sounds. Table 3.6 shows the values used for the model of the road vehicles and worked well for these types of sounds.

Window Length	Guide Ratio	Min Frequency	Max Frequency
128	0.05	10000 Hz	22050 Hz
512	0.35	4000 Hz	10000 Hz
2048	0.4	180 Hz	4000 Hz
32768	0.2	40 Hz	180 Hz

Table 3.6.: Window length, the Guide ratio and the Filter Band Limits used in the Multi-Pass Analysis.

3.5. Energy Scaling

When the final synthesis of both the tonal and noise components is complete, the two time domain audio signals can be combined by simple addition. However, during the process of analysis, modeling and synthesis, some amount of the total energy might be lost in the simplifications done. While this should not generally matter in the perception of the sound, since the SMS technique is designed to produce perceptibly similar audio, it can affect in a couple of specific instances.

While small changes in overall energy level might not be perceptible, the comparison between the levels of the tonal and the noisy components can be perceived. Such a difference is possible as very different type of algorithms and models are used for the two components and hence energy can be lost in different ways in each of the models. However a mismatch ratio or energy between the two components can be perceived making the final sound noisy or tonal.

Furthermore, the total energy content of the final synthesized signal in comparison to the original signal is important since they too are to be compared. When listened to one after each other, small variations in total energy level are audible as loudness differences in the sounds.

Hence to ensure that lost energy is compensated in an accurate way, energy based amplitude scaling was implemented in various parts of the synthesis section. Such an amplitude scaling method ensures that the total energy of the synthesized section of the sound is equal to the total energy of the corresponding section of the original sound by scaling the synthesized sound by an appropriate amount. Equation 3.17-efeqn:energyScalingE- illustrates how this can be done :

$$E_{orig} = \sum_0^N x^2 \tag{3.16}$$

$$E_{resynth} = \sum_0^N x_{resynth}^2 \tag{3.17}$$

$$x_{scaled} = x_{resynth} \times \sqrt{\frac{E_{orig}}{E_{resynth}}} \tag{3.18}$$

Here, x is the original signal, and the E_{orig} is the calculated energy term over the entire signal. $x_{resynth}$ is the synthesized signal and finally x_{scaled} is the scaled version of the synthesized signal.

This ensures that, when combined, the original and re-synthesized sounds are similar in perceived amplitude as well as in the combination of tonal and noisy components.

This scaling was used in a few parts of the synthesis including the synthesis of individual Residual hops; during the combination of synthesized noise and tonal signals; and finally overall on the final synthesized signal to make sure that the energy of the original and synthesized is the same.

3.6. Parameter Control

With the various operations in the SMS method requiring constants which change the way the analysis or the synthesis is done, the method requires a number of parameters. These parameters include, window length ,number of critical band sections, noise threshold, α for peak continuation, guide ratios, sleep interpolation width, peak energy filter minimum value, and the SavitzkyGolay filter parameters.

The control of these parameters and the interaction of their effect with one another, make a complex set of combinations that can be used for this Spectral Modeling technique. While there are definitely ways to automate the detection and calculation of optimized values of many of the parameters, most require pre-analysis of the sound and can get computationally expensive.

The current manual tuning method for the parameters is only able to tune the SMS for a specific type of sound. To be able to extend the method to various types of traffic sounds or even other source data from road vehicles, the parameters have to be re-tuned manually. Table 3.6 gives a list of the parameter values generated after tuning the SMS system for the source data used here.

3.7. Propagation Transfer Function

In a common Auralization scenario, the propagation transfer function is calculated between the source and receiver point and the corresponding time domain impulse response is convolved with the source audio to produce a representation of the audio at the receivers position.

Section	Parameter Name	Value
Multi-pass STFT Analysis	Window Length 1	128
Multi-pass STFT Analysis	Window Length 2	512
Multi-pass STFT Analysis	Window Length 3	2048
Multi-pass STFT Analysis	Window Length 4	8152
Multi-pass STFT Analysis	Filter Band Limits 1	10kHz-22kHz
Multi-pass STFT Analysis	Filter Band Limits 2	4kHz-10kHz
Multi-pass STFT Analysis	Filter Band Limits 3	180Hz-4kHz
Multi-pass STFT Analysis	Filter Band Limits 4	40Hz-180Hz
Multi-pass STFT Analysis	Hop Overlap Factor	75%
Multi-pass STFT Analysis	Zero Padding Factor	4
Peak Tracking	Noise Threshold	-65dB
Peak Tracking	Max Peaks per Frame	20
Peak Tracking	Initial Guide Number	100
Peak Tracking	Guide Increment Factor	0.25
Peak Tracking	Sections per Critical Band	25
Peak Tracking	Guide Ratio 1	0.2
Peak Tracking	Guide Ratio 2	0.4
Peak Tracking	Guide Ratio 3	0.35
Peak Tracking	Guide Ratio 4	0.05
Peak Tracking	α	0.2
Peak Tracking	Peak Energy Filter Minimum Value	0.002
Peak Tracking	Sleep Interpolation Width	0.5s
Residual Parameterization	Savitzky-Golay Pole Order	2
Residual Parameterization	Savitzky-Golay Filter Order	19

Table 3.7.: Various parameters used in the Multi-Pass Analysis.

However, with the SMS model, another approach opens up as all the information in the model is saved in the frequency domain. The transfer function, which is a scaling for each frequency can be directly applied to the model parameters. Appropriate frequency trajectories can be scaled down by the value of the propagation transfer function at that frequency, and the noise parameters can also be scaled down an averaged value of the propagation transfer function across the respective frequency band. Furthermore, since the SMS model is based on STFT based analysis, it can be used with time based propagation transfer functions, allowing various transfer functions to be applied to various time instances (or frames) of the sound model. This allows the models to be able to directly synthesize the propagated sound, removing a need to store the propaga-

Pass Number	Parameter Name	Parameter Value
Pass 1	Window Length	512
Pass 1	FFT Length	2048
Pass 1	Hop Length	128
Pass 1	Lower Frequency Limit	4000 Hz
Pass 1	Upper Frequency Limit	22050 Hz
Pass 1	Guide Ratio	0.2
Pass 2	Window Length	2048
Pass 2	FFT Length	8129
Pass 2	Hop Length	512
Pass 2	Lower Frequency Limit	180 Hz
Pass 2	Upper Frequency Limit	4000 Hz
Pass 2	Guide Ratio	0.5
Pass 3	Window Length	8192
Pass 3	FFT Length	32768
Pass 3	Hop Length	2048
Pass 3	Lower Frequency Limit	20 Hz
Pass 3	Upper Frequency Limit	180 Hz
Pass 3	Guide Ratio	0.3
All	Max number of Peaks	20
All	Noise Threshold (dB)	-65dB
All	Critical Band Sections	25
All	Guide Limit Increase Factor	0.25
All	Minimum Guide Energy Factor	0.02
All	Peak Trajectory Smoothing Factor, α	0.2

Table 3.8.: Parameters used for sounds in the listening test.

tion transfer function and calculate convolution of the impulse response and synthesized sound, which can be computationally intensive.

Care needs to be taken however though as such a use-case may come with its own set of limitation and introduce additional artifacts into the synthesis.

Combined with the propagation transfer function, an SMS based model may be used to generate real time Auralization vehicle sounds at any specified location taken much less time and also reducing memory and computational cost significantly.

3.8. Binaural Listening

For the LISTEN demonstrator, binaural renderings of the source data are needed at the receiver location based on the source model and the propagation data. While most of the binaural calculations have to do with the propagation transfer function, certain aspects of the binaural signal processing are related to the source model. In heavier vehicles and buses, when the engine is located at a different location compared to the wheels, the binaural transfer function for the engine noise and the tire noise is different. While this distinction also exists in a mono synthesis, the perceptual sensitivity of source location is greater in a binaural listening case, hence, in a binaural case is it critical to be able to separate the engine and tires noise sources, and apply separate transfer functions to them based on the locations.

Using the SMS technique the tonal component can be assumed to be the model of the engine noise and the noise component as the contribution of the tires as a rough approximation. This can enable the use of separate transfer functions to generate a binaural listening sound. However, this separation method and its effects were not studied as a part of this thesis project.

3.9. MATLAB Implementation

The entire SMS Analysis/Synthesis system was implemented in MATLAB mainly because of its ability to generate quick prototypes and also because of built-in support for many signal processing libraries and filters. The MATLAB DSP Toolbox had to be used for certain specific functions, but the rest of the system was coded in using the basic MATLAB functionality.

The MATLAB implementation allowed the visualization data, like spectrograms, magnitude spectrums and other parameters, being captured during the analysis as well as rapid testing of effects of parameter sets on the analysis and synthesis. The ability to play back the auralized model in MATLAB, helped to fix many of the issues faced in the modeling of the vehicle noise data. MATLAB scripts were also used in automating the generation of a large number of files which were used as the test data for listening

tests.

4. Analysis and Results

In this chapter the implementation of the SMS Analysis/Synthesis system done in MATLAB is analyzed for performance and the results are presented.

The aim of this auralization approach was to generate perceptually similar sounds, which are not necessarily mathematically similar. Thus, simple numerical comparisons might not yield much information of the ability of the system to model and auralize the sounds. Nonetheless, energy analysis and spectral comparisons can be used to get a basic understanding of the success of the auralization.

A perceptual test can give a better indication of the ability of the system to synthesize the sounds well. Hence, a listening test was also performed using external participants to compare sounds and gauge the performance of the auralization.

4.1. Numerical Results

4.1.1. Energy Analysis

Energy analysis, which is a form of numerical analysis, was performed at various stages of the Analysis/Synthesis system. As explained in section 2.3, the flow of energy through the various components of the model is critical in understanding the working of SMS technique. For the analysis of this method, energy analysis was performed in sections, before and after many intermediate steps, calculating the total energy before and after each step. Table 4.1 shows the energy analysis between various steps for an example analysis done on a recording of the King Long Bus driving at 90kmph. While the units of energy calculated is arbitrary using Eq. .2.20, the comparison between the values gives an idea of the flow of energy during the entire modeling process.

A few interesting patterns can be see from these values. Before the energy thresholding process, the peak trajectories track a lot more energy than there exists in the original sound. This is due to the interpolation and amplitude smoothing (sleeping and removal of small sleep durations) functionality of the trajectories. This also proves that the trajectory-tracking algorithm tends to track peaks that might not be tones but just peaks generated by higher noise energy in that frequency for that frame and assumes them as being tones. The energy threshold filter however does help to reduce a large part of these inaccurately tracked trajectories.

Another observation is that some energy seems to have been lost in the removal of the

deterministic signal from the original signal in order to generate the Residual signal. The can be attributed to the method used in spectral subtraction where the magnitude is not allowed to reduce below zero even if the deterministic spectrum has higher magnitude than the spectrum of the original signal in the same frequency bin.

Finally, the total energy level at the end of the synthesis is also about 10% less than the energy at the beginning. This difference is the most important metrics that has to be analyzed. While the nature of the modeling process definitely causes the energy to be lost in the various stages, a number which is significantly close to the original value is critical in ensuring that model was accurate in capturing the important aspects of the sound. However, differences in the energy levels do not indicate a definite perceptual difference in the sounds, but they do indicate a possibility of such a difference.

Finally, the difference in the original and re-synthesized sound energy levels means that the perceived loudness of the two sounds might be different. This has to be taken into account when designing the listening tests in Section 4.2. For the listening tests, the re-synthesized sound has to be scaled to have the same energy as the original to ensure similar loudness levels.

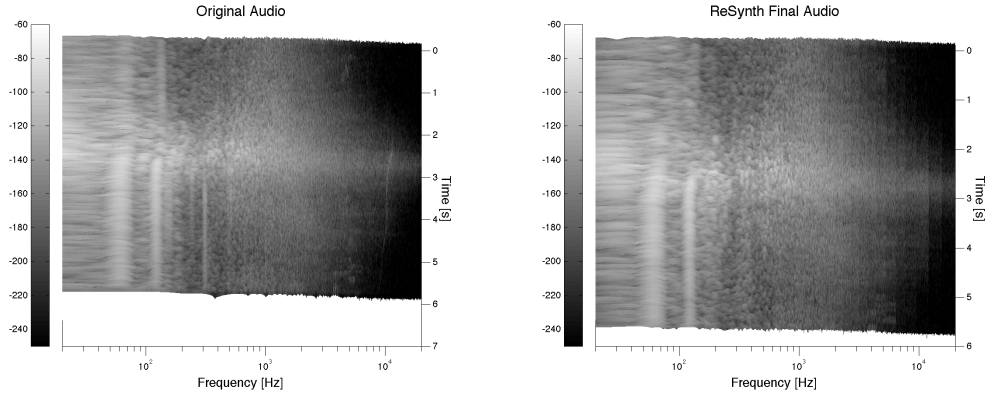
Sound	Energy [dB re-arbitary]
Original Sound	78.359
Peak Trajectories (before Energy Tresholding)	85.553
Peak Trajectories (after Energy Tresholding)	38.3
Residual	28.53
Parameterized Residual	30.102
Parameterized Residual (after Parameter Smoothing)	29.787
Re-Synthesied Tones	38.3
Re-Synthesized Noise	29.487
Re-Synthesized Combined Sound	66.548
Re-Synthesized Combined Sound (Energy Scaled)	78.359

Table 4.1.: Energy values in arbitrary units before and after each step of the SMS process normalized to the original energy.

4.1.2. Spectral Comparison

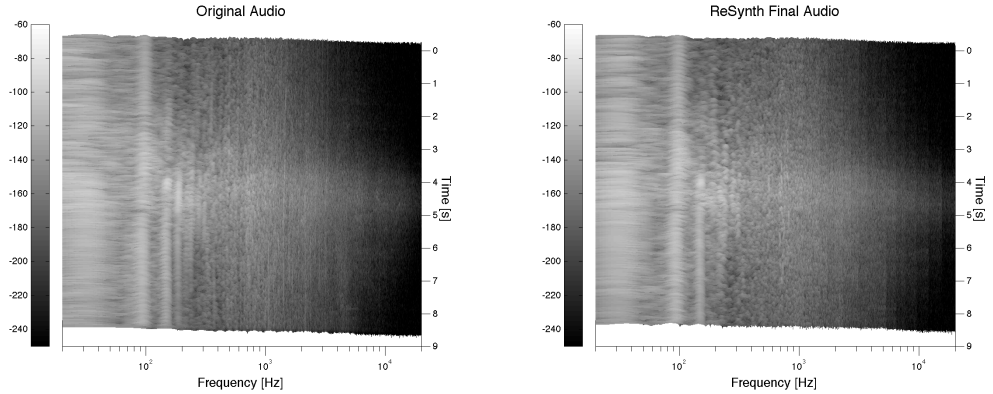
While energy analysis gives an idea of net energy levels in the various components of the modeling technique, a spectral comparison allows a visual inspection of the differences between the original sound and re-synthesized sound. The MATLAB *spectrogram* function can be used to plot spectrograms of the original sound and re-synthesized sound for comparison. Figures 4.1-4.3 shows a few of the sounds which have been re-synthesized

based on the source data.



(a) Original (b) Re-Synthesized

Figure 4.1.: Spectrogram Comparisons of a passage of a King Long bus at 90kmph.

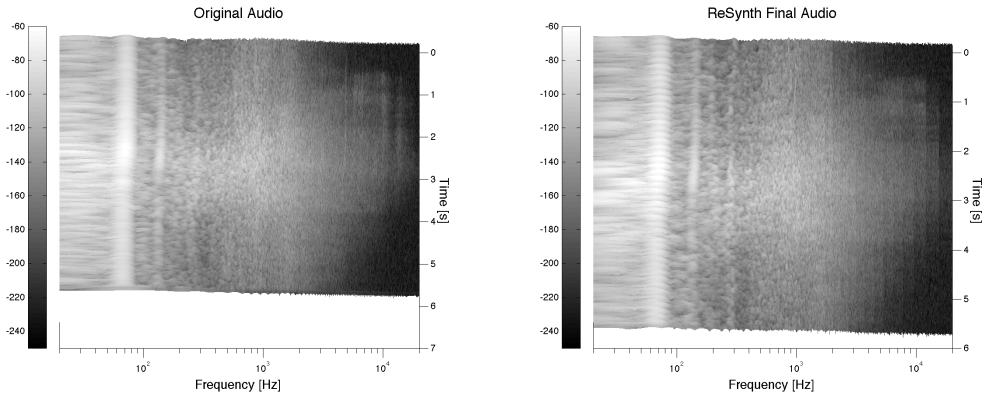


(a) Original (b) Re-Synthesized

Figure 4.2.: Spectrogram Comparisons of a passage of a King Long bus at 30kmph.

Three inferences can be drawn from the spectral comparison. A large portion of the low frequency content is modeled well by the peak tracking component of the algorithm, however, many redundant peaks are also tracked and generated into trajectories by the algorithm. These might not be audible when heard along side the other trajectories and the noise, because of their low amplitude. Hence the tracking and modeling of these peaks is redundant and should be avoided.

The parametrization of the noise into critical bands generates noise with very specific



(a) Original (b) Re-Synthesized

Figure 4.3.: Spectrogram Comparisons of a passage of a Iveco Medium Heavy at 40kmph.

band edges visible in the noise amplitude. Such edges should not affect the sound perceptually [Goodwin 96], but the bands are visible in the spectrogram. Furthermore, synthesized Gaussian noise that has much greater amplitudes than the original uniform noise, which is noticeable in the spectrogram. However, as discussed in Section 2.3.2, the amplitude does not affect the auditory perception of noise as opposed to the total energy of the noise in that frequency band, which has been kept constant in the SMS process.

The balance between the tonal component and noise component has been kept through the Analysis/Synthesis process especially with the addition of the Energy scaling as explained in Section 3.5. This is critical to have a good perception of the vehicle sound. While it is harder to see this comparison in the energy analysis, the levels of the spectrogram of the re-synthesized sound are similar to the original sound. This can be confirmed in the listening test results as well.

4.2. Listening Tests

To analyze the ability of the implemented SMS system to auralize sounds, a listening test was conducted to test the perceptual fidelity of the re-synthesized sound. The aim of the test was to know if the original and re-synthesized sounds were significantly different when heard by humans.

4.2.1. Test Design

Since the entire test revolves around comparison of the original and re-synthesized sound, a matched pair test [Orfanidis 96] was chosen to compare the two sound.

A matched pair test exposes the subjects to a sound pattern with a pair of sounds and allows them to choose one of the pair based on some criteria. The results are then tabulated to look at how many participants choose which of the two sounds in each of the pairs. It can be hypothesized that if the sounds are perceptually similar, the participants would not be able to tell the difference and hence choose both the original or re-synthesized sound with equal probability. Hence a matched pair difference test will tell us if the difference between the choice of either original or re-synthesized sound over the other is significant or not.

Selecting the original and corresponding re-synthesized sounds together and concatenating them one after each other generated the test sound patterns. A total of 18 original sounds were chosen based on the available recordings of the various vehicles. Table 4.2 shows the sounds used for the listening test.

	Filename	Vehicle Type	Speed (kmph)
1	s_otto34_v45kmph.wav	Opel Astra	45
2	vd_30_vx2_6s-trimmed.wav	Volvo V70	30
3	vd_50_vx4_6s-trimmed.wav	Volvo V70	50
4	vd_70_vx5_6s-trimmed.wav	Volvo V70	70
5	vd_90_vx5_6s-trimmed.wav	Volvo V70	90
6	iv_20_vx2_6s-trimmed.wav	Iveco Medium Heavy	20
7	iv_31_vx3_6s-trimmed.wav	Iveco Medium Heavy	31
8	iv_40_vx3_6s-trimmed.wav	Iveco Medium Heavy	40
9	iv_50_vx4_6s-trimmed.wav	Iveco Medium Heavy	50
10	iv_70_vx5_6s-trimmed.wav	Iveco Medium Heavy	70
11	buss_21_gear2_cal-trimmed.wav	King Long Bus	21
12	buss_22_gear2_cal-trimmed.wav	King Long Bus	22
13	buss_30_gear3_cal-trimmed.wav	King Long Bus	30
14	buss_39_gear4_cal-trimmed.wav	King Long Bus	39
15	buss_48_gear5_cal-trimmed.wav	King Long Bus	48
16	buss_70_gear7_cal-trimmed.wav	King Long Bus	70
17	buss_86_gear7_cal-trimmed.wav	King Long Bus	86
18	buss_91_gear8_cal-trimmed.wav	King Long Bus	91

Table 4.2.: Sounds used in the listening test.

Since the sounds were of vehicle passing by, all the sounds were edited such that the

point of passage was exactly at the midpoint of length of the audio. The sounds were also trimmed to a duration 5 s before being analyzed and synthesized. The duration of 5 s was chosen to ensure that participants were able to recall the first sound in the test while the second sound was being played to be able to compare the sounds [Nielssen 11]. As explained in Section 4.1.1 there was a difference in the total energy levels between the original and the re-synthesized sound. This was perceived as a difference in the loudness of the sounds. Hence to ensure that such factors do not affect the listening test, the re-synthesized sound was scaled to ensure that it has the same energy level as the original sound and thus a similar loudness level as described in Section 3.5.

Table A.1 in Appendix A shows the SMS parameters used for the analysis and synthesis for the listening test.

The original and re-synthesized sounds were concatenated to make a test pattern. A silence of 0.5 s was added before the first sound as well as between the two sounds to allow the participants to distinguish them. After the second sound, a pause of 4 seconds was added to ensure that the participants had enough time to answer the question.

The sounds were categorized by the four types of vehicles. Since the light vehicles have softer sounds, and the heavier vehicles have louder sounds, the patterns were played back according to categories from the light to heavy vehicles, so as to avoid the listeners ears from being unable to perceive softer sounds after hearing loud sounds.

The sound patterns were generated in both orders, with the original sound being played first (AB) and it being played second (BA). This was to reduce the effect of bias towards the play order of individual sounds in a pattern. Furthermore, each order of pattern was played twice to reduce the effect of bias towards the play order within each category. Thus, each pair was played 4 times, within a section of the listening test. Thus, 72 sound patterns were created per criteria.

Table A.2 in Appendix A shows the order of pairs of sounds played during the listening test.

Three questions were asked as a criterion for the choice between two sounds. These questions tested the three most important aspects of the vehicle sound that need to be reflected in the synthesis. The realism of the re-synthesized sounds is important to make the listener believe that they is listening to the actual vehicle. The perceived annoyance factor of the vehicles is critical as it is most often the big aspect of decision making in soundscape design and thus, has to be tested to be similar between the original and re-synthesized. Finally the perceived speed of the vehicles is an important which affects annoyance as well and hence is also a factor in the decision making process. Thus the questions listed in Table 4.3 were used in the test.

Section	Question	Number of Sound Pairs
1	Which of the two sounds is a real recording?	72
2	Which of the two sounds is more annoying?	72
3	In which sound is the vehicle moving faster?	72

Table 4.3.: Questions asked in the listening test.

4.2.2. Statistics

A test comparing two things is statistically modeled as a match pair difference t-test. This test decides of there is significant difference between the choices of either one of the sounds. In the case of the listening test, since the aim is to find if there is any difference in the choice of the sounds, regardless if its favouring the original or the re-synthesized, a two-tailed test has to be performed.

For this test, our null hypothesis, H_0, is that there is $\mu_d = 0$, where μ_d is the true difference between the number of time a sound was chosen in the specific pair by the population. And thus our alternative hypothesis, H_1, is $\mu_d \neq 0$. Thus, our null hypothesis states that the expected number of times either of the sounds is chosen, is equal. The test shall be done at 90% confidence level, which is sufficient for such an application.

4.2.3. Listening Test Implementation

The listening test was carried out between 7th September and the 17th of September 2011 at the Division of Applied Acoustics, at Chalmers University of Technology, teaching lab. The participants were tested in groups of 1 to 7 at a time. The participants listened to the list of 72 combinations of the pairs for each of the three questions and answered which of the two sounds answered the question more aptly. The answers were shaded into circles of an optical marking sheet. The participants were given a 5 minute break between each question and the corresponding set of 72 sound patterns. The participants listened to the sounds on a pair of Sennheiser HD414 open back headphones along with a sub woofer that enhanced the lower frequencies. Two separate NAD 3020 amplifiers through a M-Audio Mobile Pre USB audio interface fed both headphones and sub woofer. The headphones and the sub woofer were calibrated to give a sound level of 94 dB at the listener position for a sound being played which was -6 dB from the maximum digital sound level. A total of 28 participants were tested, out of which 21 sets of completely valid data was gathered.

4.2.4. Listening Test Results

Figure 4.2.4 shows a summary of the average results of the listening test. For each sound, the graph compares the percentage of pairs where the original sound was chosen by the participants to answer the criteria. A result of 50 % would indicate that both sounds were chosen equally often, implying the inability of the participants to make any difference between the sounds with respect to the criteria.

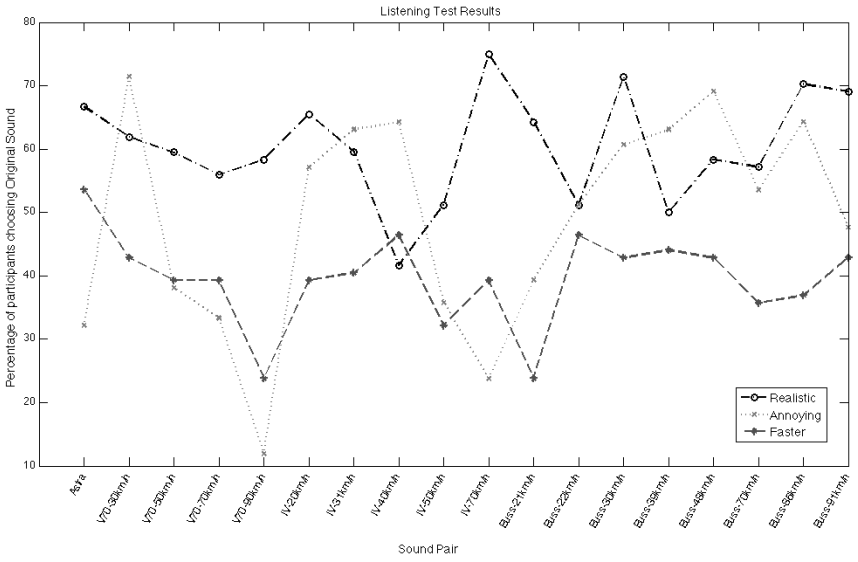


Figure 4.4.: Summary of average results from the listening test.

Considering the criteria of realism, a tendency can be observed towards the choice of the original sound, indicating a slight awareness and differentiability of the sound. Similar with speed, the tendency towards the re-synthesized sounds indicates that the perception of the speed of the vehicles in the re-synthesized sounds is greater. With annoyance, not much inferred, although it seems that the re-synthesized version of the pass-by of V70-90 was considered to be annoying by a large number of participants, possibly indicating a specific issue with the synthesis of that sound.

The plot gives a general idea of the perception of the re-synthesized sounds, but a statistical significance test has to be done to be able to claim the lack of significance in the differentiability of the two sounds.

4.2.5. Statistical t-test

Based on the matched pair t-test setup as described in Section 4.2.2, the test statistic for each pair can be calculated. An example for a specific sound (King Long Bus traveling at 70kmph) is shown in Table 4.4.

Listener	Original	Re-Synthesized	Difference, d	$(d-\overline{d})^2$
1	1	3	-2	6.61
2	0	4	-4	20.90
3	4	0	4	11.76
4	4	0	4	11.76
5	3	1	2	2.04
6	2	2	0	0.33
7	1	3	-2	6.61
8	4	0	4	11.76
9	2	2	0	0.33
10	2	2	0	0.33
11	3	1	2	2.04
12	1	3	-2	6.61
13	3	1	2	2.04
14	4	0	4	11.76
15	3	1	2	2.04
16	3	1	2	2.04
17	1	3	-2	6.61
18	2	2	0	0.33
19	4	0	4	11.76
20	0	4	-4	20.90
21	1	3	-2	6.61

Table 4.4.: Calculation for T-Test for the sound pair of the King Long Bus traveling at 70kmph for the question of realism of the sound.

Firstly, a new variable, d can be defined to indicate the difference between the paired values. In this case, it would be the difference between the number of times a original or re-synthesized sound was choosen. Eq. 4.1 defines this,

$$d_i = n_{i,orignal} - n_{i,re-synthesized} \tag{4.1}$$

where $n_{i,orignal}$ is the number of times the original version of a specific sound was chosen by the i^{th} participant and $n_{i,re-synthesized}$ is the number of times a specific sound was chosen by the i^{th} participant.

Next, the sample mean, $\overline{d}$ of the differences is calculated as per Eq. 4.2,

$$\overline{d} = frac(\sum_i d_i)n \tag{4.2}$$

where n is the total number of participants.

Next, the standard deviation of the sample is calculated based on Eq. 4.3:

$$s_d = \sqrt{\frac{\sum(d_i - \overline{d})^2}{(n-1)}} \tag{4.3}$$

Based on the standard deviation, a standard error of the entire population can be estimated. In the case of the population, N, being much larger ($N >> n$) than the sample size, n, Eq. 4.4 can be used as an estimate of the standard error :

$$SE = \frac{s_d}{\sqrt{(n)}} \tag{4.4}$$

Hence the test statistic can be expressed as:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - D}{SE} = \frac{\overline{d} - D}{SE} \tag{4.5}$$

where D is the hypothesized mean difference between the population pairs, which in the case of the null hypothesis is 0.

Looking up the test statistic at significance level in the t-distribution gives p-value of the t-statistic. If the p-value is greater than the significance level, then it can be concluded that the null hypothesis cannot be rejected.

MATLAB also provides a method, *ttest*, which automates the calculation of the test results. Running it on the results from the listening test gives the acceptance or rejection of the null hypothesis. In the cases of the acceptance, a conclusion can be drawn that for that specific criteria, there is no significant difference in the original and re-synthesized sounds. For the rejected cases, the conclusion has to be drawn that there is significant difference that the participants are able to detect. Table 4.5 tabulates the results of the t-test. Overall, the null hypothesis had to be rejected 30 times, while it was accepted 24 times.

Pair	Vehicle	Speed (kmph)	Realism	Annoyance	Speed
1	Opel Astra	45	Rejected	Rejected	Accepted
2	Volvo V70	30	Rejected	Rejected	Accepted
3	Volvo V70	50	Rejected	Rejected	Rejected
4	Volvo V70	70	Accepted	Rejected	Accepted
5	Volvo V70	90	Accepted	Rejected	Rejected
6	Medium Heavy	20	Rejected	Accepted	Rejected
7	Medium Heavy	31	Accepted	Rejected	Rejected
8	Medium Heavy	40	Accepted	Rejected	Accepted
9	Medium Heavy	50	Accepted	Rejected	Rejected
10	Medium Heavy	70	Rejected	Rejected	Accepted
11	Bus	21	Rejected	Accepted	Rejected
12	Bus	22	Accepted	Accepted	Accepted
13	Bus	30	Rejected	Rejected	Accepted
14	Bus	39	Accepted	Rejected	Accepted
15	Bus	48	Accepted	Rejected	Accepted
16	Bus	70	Accepted	Accepted	Rejected
17	Bus	86	Rejected	Rejected	Rejected
18	Bus	91	Rejected	Accepted	Accepted

Table 4.5.: Results of the t-test for significant difference on the listening test sounds.

5. Discussion

This chapter looks at implementation of the Spectral Modeling Synthesis technique on traffic noise and the results of the listening test. The implementation is compared with the original goals of the project. Implications of the results are discussed as well as the key findings and lessons learned in the process of this project work.

5.1. SMS Technique

At the most fundamental level, an important result of this thesis project is the ability of the SMS technique to be applied too much nosier sounds like traffic noise. While the SMS technique was developed for modeling musical instruments and sounds, its direct application with minimal changes to traffic noise is a significant result. While not specifically designed to model the noisy component of a given sound with complex algorithms, the SMS technique is able to capture and reproduce the essence of the noisy traffic sounds accurately using enveloped white noise.

Furthermore, the concept of separation of tonal and noisy components, and the ability of including the propagation effects easily and without much computation into the synthesis model, lends itself well to traffic noise model, where the vehicle engines and tires generate significantly more tonal and noisy sounds respectively. This also opens up potential for using the parameters of the tonal and noisy components to synthesize various types of sounds.

5.2. Source

The source material used in the analysis stage is very critical in such a model. The source recording has to capture the essence of the vehicle sound, and not capture other ambient sounds. While some ambient sounds were removed during the pre-processing of the recording (see Section 3.1), some other sounds (for e.g. birds chirping) were too well integrated into the recording to be able to be removed by post processing. Since the analysis parameters were not tuned to capture such sounds and they noticeably lacking in the re-synthesis.

After the listening test, few participants pointed out that certain un-natural sounding artifact in the synthesized caused the re-synthesized sound to be detectable with high

accuracy and repeatability. These could also be due to non-traffic noise being present in the source material.

The type of vehicles analyzed is also something to be considered. Much of the source material used in this thesis work was based on the availability of good clean recordings. While that allowed the testing of this technique on various types of vehicles (small, medium and large), for a better model, trying out the model with a wider variety of vehicles would allow a more generic model to be developed.

5.3. Adjustments and Changes

The improvements done to the SMS Algorithm were intended to improve the perceptual accuracy of the model as well as reduce the time taken for the analysis process. While in some cases it was a trade off between the accuracy and speed, manual adjustments and tweaking allowed for a viable trade-off. The extra information captured by the multiple-pass analysis and the selective discarding of information in the peak-tracking algorithm helped to improve the perceptible accuracy of the models, while still being able to generate the synthesis in a realistic amount of time.

A point to be noted here is the software model developed for ease of analysis and the ability to easily plot, compare and manipulate data and parameters. Hence, in many occasions computational efficiency was not given a priority over development convenience leading to a generally longer amount of time taken to analyze and synthesize.

While there was no listening test or numerical analysis done to compare the actual effect of all the additions and improvements to the original SMS technique, a large part of the perceptual accuracy of the synthesized sound was a result of the various additions and improvements. Furthermore, many of the adjustments and changes were specifically done to reduce certain effects or artifacts (see Section 3.2.4).

5.3.1. Analysis Parameter

It was noticed during the development of the technique that the accuracy of the system was extremely sensitive to analysis parameters. Small changes in analysis parameters would affect the sound significantly. While no data was collected on this, a large number of adjustments had to be done to improve the perceptual accuracy of the synthesis. This was a critical and affected the modeling significantly even when the adjustment itself was small.

The sensitivity of Analysis parameters implied a significant importance on some technique to be able to control or tune these parameters, especially for the specific source sounds being used. Manual tuning, which was used widely throughout the thesis work can yield some good results. This was visible in the listening test results where specific sounds (Pair 4 - Volvo V70 at 70kmph, Pair 12 - King Long Bus at 22kmph, Pair 14 -

King Long Bus at 39kmph), which were used in development, had very little significant difference between the original and the synthesized sound, which meant they were very similar. This could indicate that the analysis parameter were better tuned for those specific sounds and hence worked better with them.

Manually tuning the analysis parameters for each individual of the 18 source files would have probably yielded better-synthesized sounds, but that would be very time consuming. Hence, a more automated system for doing such tuning would be useful in such a situation. None the less, this does indicate that a well-tuned system would be able to yield significantly similar sounds.

5.3.2. Energy Analysis

Energy Analysis turned out to be a very useful tool during this thesis work. The versatility of energy analysis and its ability to detect loss of information was invaluable in both development as well as debugging of the technique. Energy analysis ensured that the loudness of the individual components as well as of the combined sound remained similar after the analysis/re-synthesis. This too was critical, in the light of a matched pair test where a small difference in loudness would be easily detectable.

5.4. Listening Tests

The listening tests done with the adjustments and improvements to the SMS Technique did not prove statistically significant lack of difference between the original and synthesized sounds across all of the sources used, when tested using a t-test. A number of the t-tests had to be rejected for the alternative hypothesis that the sounds were noticeably different

With reference to Table 4.5 and especially Figure 4.2.4, some vague trends can be seen where the results were closer to the 50% mark, which would then imply a lack of noticeable difference. The sounds of the King Long Bus tended to be closer to the 50% mark on more occasions and also tended to be accepted more often than the other types of vehicles in the t-test as seen in Table 5.1.

One source file, Pair 12 - King Long Bus at 22kmph, shows exceptionally good performance and has all three of the t-tests accepted. As discussed in Section 5.3.1, this could be a result of a good tuning of the analysis parameters for that source file.

Similarly, the question on Realism and Speed had many more accepted t-tests than the question on Annoyance on an aggregate scale as seen in Table 5.2. Hence the Annoyance seemed be a characteristic where the listeners had a more polarized opinion about the sounds. This can be an indication that the SMS technique is changing the sounds making them significantly more or less annoying (depending on the source) than than the original.

Vehicle	Total T-Tests	Accepted T-Tests	Acceptance Ratio
Opel Astra	3	1	0.33
Volvo V70	12	4	0.33
Medium Heavy	15	6	0.4
King Long Bus	24	13	0.55

Table 5.1.: Tabulation for aggregated T-Test results for each vehicle type

Question	Total T-Tests	Accepted T-Tests	Acceptance Ratio
Realism	18	9	0.50
Annoyance	18	5	0.29
Speed	18	10	0.55

Table 5.2.: Tabulation for aggregated T-Test results for each question type

Furthermore, it can be argued that a t-test is a very strict measure of the modeling accuracy of the technique. While the aim of the model is to generate a perceptually similar sound, a matched pair test, specifically tests for equivalence of the sound. In many cases, including the target scenario of the urban environment simulator, it is sufficient for the synthesized sound to be similar to the original, and does not have to be exactly equal to the original. Furthermore, in the target scenario, the synthesized sounds would go through further processing based on propagation effects before being heard by the listener, possibly increasing the realism of the sounds.

A listening test which combines the source models generated using this technique with propagation effects would yield more accurate information about the validity of this technique in the simulator.

References

[Gabor 46] Gabor D. : *Theory of Communications*, Journal of the IEEE, Vol. 93, 1946.

[Kleiner 93] Kleiner M., Dalenbäck B. and Svensson P.: *Auralization - An Overview*, Journal of Audio Engineering Society, Vol 41, No 11, 1993

[Forssén 09] Forssén J., Kaczmarek, T., Alvarsson, J., Lundén, P. and Nilsson, M.: *Auralization of traffic noise within the LISTEN project preliminary results for passenger car pass-by*, Euronoise 2009, pp. 26-28, 2009

[Plovsing 06] Plovsing B., Kragh J.: *Nord2000. comprehensive outdoor sound propagation model. Part 1: Propagation in atmosphere without significant refraction*, Delta Acoustics & Vibration Report AV 1849/00, 2006

[van Maercke 07] van Maercke D. and Defrance J.: *Development of an analytical model for outdoor sound propagation within the harmonoise project*, Acta Acustica united with Acustica, vol. 93, pp. 201212, 2007

[van der Aa 10] van der Aa, B.: *Ground buried resonators - Analytical and numerical modeling of their noise reducing effect for sound propagating outdoors from traffic noise sources*, Thesis work, Department of Applied Acoustics, Chalmers University of Technology, Göteborg, 2010

[Chowning 73] Chowning J.M.: *The Synthesis of Complex Audio Spectra by Means of Frequency Modulation*, Journal of Audio Engineering Society, Vol 21, No 7, 1973

[Roads 93] Roads C.: *Introduction to Granular Synthesis*, Computer Music Journal, Vol 12, No 2, 1988

[Serra 90] Serra X. and Smith J.: *Spectral Modeling Synthesis: A Sound Synthesis/-Analysis Technique Based on Deterministic and Stochastic Decomposition*, Computer Music Journal, Vol 14, No 4, 1990

[Serra 03] Serra X.: *Spectral Modeling Synthesis: Past and Present*, Presentation at DAFx Conference, 2003

[Desainte-Catherine 00] Desainte-Catherine M. and Marchand S.: *High-Precision Fourier Analysis of Sounds Using Signal Derivatives*, Journal of Audio Engineering Society, Vol 48, No 7/8, 2000

[Risset 85] Risset J.-C.: *Computer music experiments, 1964–*, Computer Music Journal, Vol 9, No 1, 1985

[Zwicker 90] Zwicker, E. and Fastl, H.: *Psychoacoustics, Facts and Models*, Springer-Verlag, 1990

[Goodwin 96] Goodwin, E. : *Residual modeling in music analysis-synthesis*, Conference Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol 2, 1996

[Smith 12] Smith, J.O. *"Hamming Window"*, Spectral Audio Signal Processing online book, accessed 12/11/2012, `https://ccrma.stanford.edu/~jos/sasp/Choice_WOLA_Window.html`

[Markel 96] Markel J.D. and Gray A.H. : *Linear Prediction of Speech*, Springer-Verlag, 1976

[Sethares 07] Sethares W.A. : *Rhythms and Transforms* , Springer, 2007

[Orfanidis 96] Orfanidis, S.J. : *Introduction to Signal Processing* , Prentice-Hall, 1996

[Orfanidis 96] Kreyzig, E. : *Introductory Mathematical Statistics* , John Wiley, 1970

[Strang 93] Strang, G. : *Wavelet Transforms Versus Fourier Transforms* , Bulletin of the Society of American Mathematical Society, 1993

[ISO 362-1 07] ISO 362-1:2007: *Measurement of noise emitted by accelerating road vehicles – Engineering method*, International Organization for Standardization, `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=42210`

[Nielssen 11] Nielssen, M.: *Settup for Listening Test*, Email Conversation on listening test setup for matched pair test, 05 September 2011

A. Listening Tests

A.1. Listening Test Design

Parameter Name	Parameter Value
Pass 1 Window Length	512
Pass 1 FFT Length	2048
Pass 1 Hop Length	128
Pass 1 Lower Frequency Limit	4000 Hz
Pass 1 Upper Frequency Limit	22050 Hz
Pass 1 Guide Ratio	0.2
Pass 2 Window Length	2048
Pass 2 FFT Length	8129
Pass 2 Hop Length	512
Pass 2 Lower Frequency Limit	180 Hz
Pass 2 Upper Frequency Limit	4000 Hz
Pass 2 Guide Ratio	0.5
Pass 3 Window Length	8192
Pass 3 FFT Length	32768
Pass 3 Hop Length	2048
Pass 3 Lower Frequency Limit	20 Hz
Pass 3 Upper Frequency Limit	180 Hz
Pass 3 Guide Ratio	0.3
Max number of Peaks	20
Noise Threshold (dB)	-65dB
Critical Band Sections	25
Guide Limit Increase Factor	0.25
Minimum Guide Energy Factor	0.02
Peak Trajectory Smoothing Factor, α	0.2

Table A.1.: Parameters used for sounds in the listening test

No.	Vehicle Type	Speed (kmph)	Order	No.	Vehicle Type	Speed (kmph)	Order
1	Opel Astra	45	BA	37	Iveco Daily	31	BA
2	Opel Astra	45	AB	38	Iveco Daily	50	AB
3	Opel Astra	45	AB	39	Iveco Daily	50	BA
4	Opel Astra	45	BA	40	Iveco Daily	70	BA
5	Volvo V70	90	BA	41	King Long Bus	30	AB
6	Volvo V70	30	AB	42	King Long Bus	30	BA
7	Volvo V70	70	AB	43	King Long Bus	39	BA
8	Volvo V70	90	AB	44	King Long Bus	91	BA
9	Volvo V70	30	AB	45	King Long Bus	70	BA
10	Volvo V70	30	BA	46	King Long Bus	30	AB
11	Volvo V70	50	BA	47	King Long Bus	48	BA
12	Volvo V70	50	AB	48	King Long Bus	21	AB
13	Volvo V70	50	BA	49	King Long Bus	48	BA
14	Volvo V70	70	BA	50	King Long Bus	22	BA
15	Volvo V70	70	BA	51	King Long Bus	70	AB
16	Volvo V70	30	BA	52	King Long Bus	21	BA
17	Volvo V70	50	AB	53	King Long Bus	48	AB
18	Volvo V70	90	BA	54	King Long Bus	22	AB
19	Volvo V70	90	AB	55	King Long Bus	86	BA
20	Volvo V70	70	AB	56	King Long Bus	48	AB
21	Iveco Daily	70	AB	57	King Long Bus	86	AB
22	Iveco Daily	70	BA	58	King Long Bus	22	AB
23	Iveco Daily	40	BA	59	King Long Bus	21	AB
24	Iveco Daily	20	AB	60	King Long Bus	86	BA
25	Iveco Daily	40	BA	61	King Long Bus	22	BA
26	Iveco Daily	40	AB	62	King Long Bus	30	BA
27	Iveco Daily	40	AB	63	King Long Bus	39	AB
28	Iveco Daily	31	AB	64	King Long Bus	86	AB
29	Iveco Daily	50	BA	65	King Long Bus	91	AB
30	Iveco Daily	50	AB	66	King Long Bus	39	BA
31	Iveco Daily	31	AB	67	King Long Bus	91	BA
32	Iveco Daily	20	BA	68	King Long Bus	21	BA
33	Iveco Daily	70	AB	69	King Long Bus	39	AB
34	Iveco Daily	20	BA	70	King Long Bus	70	BA
35	Iveco Daily	20	AB	71	King Long Bus	70	AB
36	Iveco Daily	31	BA	72	King Long Bus	91	BA

Table A.2.: Order of sounds played in the listening test