Modelling and transformation of sound textures and environmental sounds

Wei-Hsiang Liao

Université Pierre et Marie Curie

Doctoral Thesis

Modelling and transformation of sound textures and environmental sounds

Author:
Wei-Hsiang Liao

Supervisor:
Axel Roebel
Wen-Yu Su

A thesis submitted in fulfilment of the requirements
for the degree of DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

in the

Équipe Analyse/Synthése
Institut de Recherche et Coordination Acoustique/Musique
École doctorale Informatique, Télécommunications et Électronique (Paris)

Jury:

Mr. Shlomo Dubnov	Professor, UCSD, U.S.A.	Reviewer
Mr. Josh McDermott	Professor, MIT, U.S.A.	Reviewer
Mr. Laurent Daudet	Professor, University Paris 7, France	Examiner
Mr. Bruno Gas	Professor, ISIR University Paris 6, France	Examiner
Mr. Alvin W.-Y. Su	Professor, National Cheng Kung University, Taiwan	Examiner
Mr. Axel Roebel	HDR. IRCAM, France	Examiner

Date of the Defense: July 15, 2015

Declaration of Authorship

I, Wei-Hsiang Liao, declare that this thesis titled, 'Modelling and transformation of sound textures and environmental sounds' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

UNIVERSITÉ PIERRE ET MARIE CURIE

Abstract

Institut de Recherche et Coordination Acoustique/Musique
École doctorale Informatique, Télécommunications et Électronique (Paris)
DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Modelling and transformation of sound textures and environmental sounds

by Wei-Hsiang Liao

Recently, the processing of environmental sounds has become an important topic in various areas. Environmental sounds are mostly constituted of a kind of sounds called sound textures. Sound textures are usually non-sinusoidal, noisy and stochastic. Several researches have stated that human recognizes sound textures with time-averaged statistics that characterizing the envelopes of auditory critical bands. This suggests that these statistics should be preserved while synthesizing sound textures. Existing synthesis algorithms can impose some statistical properties to a certain extent, but most of them are excessively computational intensive. In this thesis, we propose a new analysis-synthesis framework that contains a statistical description that consists of perceptually important statistics and an efficient mechanism to adapt statistics in the time-frequency domain. The quality of resynthesised sound is at least as good as state-of-the-art but more efficient in terms of computation time. The statistic description is based on the short-time-Fourier-transform. However, if certain conditions are met, the proposed mechanism can also adapt to other filter bank based time-frequency representations. The adaptation of statistics is achieved by utilizing the connection between the statistics on time-frequency representation and the spectra of time-frequency domain coefficients. If the order of statistics is not greater than two, feasible signals can directly be generated from statistical descriptions without iterative steps. When the order of statistics is greater than two, the algorithm can still adapt all the statistics within a reasonable amount of iteration. It is possible to adapt only a part of cross-correlation functions. This allows the synthesis process to focus on more important statistics and ignore the irrelevant parts, which provides extra flexibility. With the proposed framework, one can easily extract the statistical description of a sound texture then resynthesizes arbitrary long samples of the original sound texture from the statistical description. The proposed algorithm has several perspectives. It could possibly be used to generate unseen sound textures from artificially created statistical descriptions. It could also serve as a basis for transformations like stretching or morphing. One could also expect to use the model to explore semantic control of sound textures. . .

Acknowledgements

The accomplishment of this thesis is impossible without the support from many people, both in France and Taiwan. Therefore, I would like to represent my gratitude to whoever helped me here.

First of all, I would like to thank my thesis supervisor, Dr. Axel Roebel, for his kindness and brilliant guidance. Each time when I encounter difficulties, his knowledge and patience helped me to surpass countless obstacles. His is always willing to spend hours to discuss issues, solve problems and help revising articles. From him, I learned various practical skills in the scientific research.

I owe my sincere thanks to my co-supervisor Prof. Alvin Wen-Yu Su and Prof. Xavier Rodet. Without their help, I would not had the chance to start an international co-supervised Ph.D. program. Prof. Alvin Wen-Yu Su is kind and generous, he always provide strong supports whenever a difficulty is encountered.

The people in IRCAM are friendly and excelled in their working field. Working in IRCAM was an amazing experience to me. I would like to express my thanks to Sean O'Leáry, for his advices on sound textures, Nicolas Misdariis, for his suggestions for the perceptual test and Frédéric Cornu, for his impeccable programming advices.

It is also fortunate for me that I have many friends in Paris. Thanks to Mei-Hua, for her care and generosity. She is always glad to provide assistance. The life in Paris would not have been so fascinating without Chung-Hsin, Chia-Ling and Yi-Wen.

At last, I would like to represent my deepest gratitude to my family for their unwavering support.

Contents

List of Figures

List of Tables

Abbreviations

ACF	**A**uto-**C**orrelation**F**unctions
CCF	**C**ross-**C**orrelation**F**unctions
CF	**C**orrelation**F**unctions
CQT	**C**onstant-**Q** **T**ransform
ERB	**E**quivalent **R**ectangular **B**andwidth
MUSHRA	**MU**ltiple **S**timuli with **H**idden **R**eference and **A**nchor
PCCFI	**P**artial **C**ross-**C**orrelation **F**unction **I**mposition
SNR	**S**ignal-to-**N**oise **R**atio
STFT	**Short**-**T**ime **F**ourier **T**ransform
TFR	**T**ime **F**requency **R**epresentation

Symbols

$x(t)$		Time-domain signal
$w(t)$		Window function
$\Re\{x\}/\Im\{x\}$		Real / Imaginary part of x
N_x		Length of x
N_k		Fourier transform size
$\mathcal{F}\{x\}_k$,$\hat{x}(k)$	$\sum_{t=0}^{N_x-1} x(t)e^{-2\pi j\frac{kt}{N_x}}$	Fourier transform of $x(t)$
$\mathcal{F}^{-1}\{y\}_t$,$\check{y}(t)$	$\frac{1}{N_y}\sum_{k=0}^{N_k-1} y(k)e^{2\pi j\frac{kt}{N_y}}$	Inverse Fourier transform of $y(k)$
$\theta(x)$	$\Im(\log x)$	Principal arg of x, $x \in \mathbb{C}$
c		Amplitude compression coefficient, $c = 0.15$
l		Hop-size of STFT
$X(n,k)$	$\sum_{t=-\infty}^{\infty} x(t)w(t-nl)e^{-2\pi j\frac{kt}{N_k}}$	The Short-Time Fourier Transform of $x(t)$
$\mathcal{X}_k(n)$	$\lvert X(n,k)\rvert^{2c}$	The sub-band signal of kth frequency bin of X
$A_x(\tau)$	$A_x(\tau) = \sum_{t=-\infty}^{\infty} x(t)\overline{x}(t+\tau)$	Auto-correlation function of $x(t)$
$C_{x,y}(\tau)$	$C_{x,y}(\tau) = \sum_{t=-\infty}^{\infty} x(t)\overline{y}(t+\tau)$	Cross-correlation function between $x(t)$ and $y(t)$
$\mu(X)$	$E[X]$	Mean of X
$\mu_n(X)$	$E[(X-E[X])^n]$	The nth central moment of X
$\sigma^2(X)$	$E[(X-\mu(X))^2]$	Variance of X
$\eta(X)$	$E[(\frac{X-\mu(X)}{\sigma(X)})^3] = \frac{\mu_3(X)}{\sigma^3(X)}$	Skewness of X
$\kappa(X)$	$E[(\frac{X-\mu(X)}{\sigma(X)})^4] = \frac{\mu_4(X)}{\sigma^4(X)}$	Kurtosis of $x(t)$

$\mathbf{X}$	$\mathbf{X} = \{\mathcal{X}_0, \mathcal{X}_1, \cdots, \mathcal{X}_K\}^T$	Matrix of sub-band signal, $K = \lfloor N/2 \rfloor$
$(\mathbf{X})_i$	$(\mathbf{X})_i = \mathcal{X}_i$	The ith row vector of matrix $\mathbf{X}$
$(\mathbf{X})_{i,j}$	$(\mathbf{X})_{i,j} = \mathcal{X}_{i,j}$	The element at ith row and jth column of $\mathbf{X}$
$\mathbf{X} \circ \mathbf{Y}$	$(\mathbf{X}_{i,j}\mathbf{Y}_{i,j}) \quad \forall(i,j)$	The Hadamard product of matrix $\mathbf{X}$ and $\mathbf{Y}$
$\hat{\mathbf{X}}$	$\hat{\mathbf{X}} = \{\hat{\mathcal{X}}_0, \hat{\mathcal{X}}_1, \cdots, \hat{\mathcal{X}}_K\}^T$	Row-wise Fourier transform of $\mathbf{X}$
$\eta(\mathbf{X})$	$\eta(\mathbf{X}) = \{\eta(\mathcal{X}_0), \cdots, \eta(\mathcal{X}_K)\}^T$	Row-wise skewness of $\mathbf{X}$
$\kappa(\mathbf{X})$	$\kappa(\mathbf{X}) = \{\kappa(\mathcal{X}_0), \cdots, \kappa(\mathcal{X}_K)\}^T$	Row-wise kurtosis of $\mathbf{X}$

To my family and my friends ...

Overview

The main goal this thesis is to seek an analysis-synthesis mechanism which can handle sound textures with more stable and efficient performance compared to the state-of-the-art. It should be as efficient as possible such that the currently prohibitively long runtime of the only existing algorithm for imposing statistical properties is reduced. The results obtained may allow efficient synthesis and transformation of sound textures and noises, and on the other hand may lead to improved algorithms that allow transformation of the aperiodic or noise components in music and speech signals.

In this article, we introduce a new sound texture analysis-synthesis framework that contains a statistical description that consists of perceptually important statistics and an efficient mechanism to adapt statistics in the time-frequency domain. The quality of resynthesised sound is at least as good as state-of-the-art but more efficient in terms of computation time. The model with statistic description is based on the short-time-Fourier-transform. However, if conditions are met, it can also adapt to other filter bank based time-frequency representations. It characterizes sound textures based on their joint statistics on time-frequency representations, forming statistical descriptions. From the statistical descriptions, the signals which fit all second order joint statistics are derived then the remaining higher order statistics are applied. This is capable of generating arbitrary many samples from a sample of sound texture. Compared to the previous works, there are several advantages using the proposed algorithm. First, it considered spectro-temporal correlations, which are dependencies of all directions on the time-frequency domain. Second, if the signal can be defined by only second order statistics, the target signal can be generated without iterations. If this is the case, statistics of the generated signals will be the exact values as described by the second order statistics. Third, the algorithm allows applying only a part of the statistics, giving more flexibility. Both objective and subjective evaluations are conducted to investigate the quality and alikeness of sounds generated from the proposed algorithm. We expect the proposed mechanism could serve as a basis for further transformations of sound textures.

The thesis is organized as follows:

Chapter 1: This introduction describes the general idea and difficulty of sound texture processing and the motivation of using statistics of the time-frequency representation. Then, perceptually important statistics of sound textures are described. The section also explains why these statistics are perceptually important. After that, the relation between the sound texture and random process are discussed. The final discussion explains the goal of and the organization of this thesis.

Chapter 2: The state-of-the-art chapter introduces the history and the previous works on sound texture synthesis. We start with early attempts in both sound and visual textures and introduce the idea of Julesz conjecture. Following sections survey different synthesis approaches, including some of the most important works of sound texture synthesis. Model-based synthesis and granular-based synthesis algorithms are both discussed in their own subsection. A final discussion section briefly explains the reasoning why the proposed approach is developed this way and talk about the relation with the previous works.

The following chapters describe the proposed method.

Chapter 3: This chapter discusses the selection of usable filter bank based TFRs. The subsections explain what the conditions are for a suitable TFR. If a TFR meets the conditions, then the proposed statistical description can be applied on the TFR. Following the discussion of TFRs, the overview of the statistical description for sound texture is given. It also explained several problems. For example, the phase coherency problem and the down-sampling effect related to the hop size used in STFT. In the end of the chapter, it describes how the statistical description actually obtained from a sound texture.

Chapter 4: In this chapter, we introduce the methods that are used to impose statistical properties. There are two methods about imposing moments to a TFR, while another two methods impose correlation functions to the TFR. The first method imposes all the correlation functions to the TFR, without using iterative approaches. This method represents a signal by its statistical properties. The second and the third method are to impose skewness and kurtosis to frequency bands. The second method can impose the exact value of the designated moments, but will change the correlation functions in the process. The third method can only imposes an approximated value of the designated moments, but preserves the correlation functions. This useful property helps to achieve an efficient imposition process. The last method gives the freedom to apply a selected part of cross-correlation functions, which provides much flexibility. In the end, the advantages and disadvantages of each method are discussed.

Chapter 5: The chapter summarizes the content of chapter 3 and 4, and gives a complete description of the proposed analysis-synthesis scheme of sound texture. Implementation details and parameter settings are also discussed in the chapter.

Chapter 6: The efficiency of the proposed algorithm is examined by profiling. Then the statistics of the resynthesized signal are compared with those in the statistical description. In the following are three perceptual tests. One of the tests investigates how the length of imposed cross-correlation function will affect to the auditory system. Other experiments compares the proposed algorithm with Bruna's(Bruna and Mallat, 2013) and McDermott's(McDermott and Simoncelli, 2011a) work. The discussion section gives a summary of discoveries that have been obtained from these experiments.

Chapter 7: The chapter briefly talks about the contributions and the discoveries in each chapter and discusses the difficulties of the algorithm then gives the conclusion. The perspectives discuss the possible future improvements of the algorithm and the potential application of the proposed analysis-synthesis framework.

Chapter 1

Introduction

1.1 What are Environmental Sounds and Sound Textures

Environmental sounds include all the sounds coming from our surroundings. Some of them are coming from natural phenomenon, such as wind blowing, rain falling or water streaming. Some of them are coming from mechanical events, such as the buzzing of motor, spinning of helicopter or the traffic noise during the rushing hour. A large portion of environmental sounds are usually regarded as **sound textures**.

These sounds are so common in our daily life that we cannot ignore their presence. In contrast to usual instrumental sounds, many of these sound textures have no sinusoidal components, no harmonic structures and no significant pitches. While some parts of the sound textures are noise-like, they still contain various structures which are induced by physical processes and randomness. (Strobl et al., 2006) categorizes common sound textures into 5 categories:

Natural sounds fire, water (rain, waterfall, ocean) wind

Animal sounds sea gulls, crickets, humming

Human utterances babble, chatter

Machine sounds buzz, whir, hammer, grumble, drone, traffic

Activity sounds chip, sweep, rustle, typing, scroop, rasp, crumple, clap, rub, walking

Although the term 'sound texture' has been widely used to describe non-instrumental sounds, there is no widely accepted definition of sound textures(McDermott and Simoncelli, 2011a). Arnaud(Saint-Arnaud and Popat, 1998) proposed an aspect that, sound

textures are composed of various kinds of atomic sound events, but that atomic sound events do not necessarily form a sound texture. For example, the sound of a raindrop falling and a clap of hand are unlikely to be called sound textures, but heavy rain falling and the applause of audience in a stadium are widely accepted as sound textures. He also mentioned that, there is no consensus among people as to what a sound texture might be; more people will accept sounds that fit a more restrictive definition. As a result, it is rather important to clarify what the 'sound texture' does mean in this thesis. Considering the properties of widely accepted sound textures, we found that there is something in common. First, sound textures sound differently in every moment of its whole duration, but human usually recognize the entire sound as a continuous stationary entity. For example, the sound of steady wind blowing, heavy rain falling and water boiling. In other words, when a human listens to several segments of a texture, the human will agree that all these segments belong to the same sound texture. Second, the human recognition of a sound texture seems do not rely on individual sound events, which is very different from how human recognize ordinary musical sounds. For example, humans usually recognize musical pieces by its melody, which is composed by a series of musical notes. In this case, the musical notes are the sound events. Humans memorize the sequence of musical notes and recognize the musical piece when the sequence is identified. However, humans do not recognize sound textures by identifying a specific sequence of sound events. Humans will not discriminate two segments of fire crackling sound by the sequences of cracklings. Following these two observations, we can see that, humans recognize sound textures according to some characteristics that spread across the entire sound segment. Therefore, we can come up with an idea of sound texture: *A sound is a sound texture if human recognizes it through something else than time-limited characteristics and these characteristics must be the same for all segments of the same sound.* This will be further discussed in later sections (chapter 3).

1.2 Research Motivation

Compared to the instrumental sounds, sound textures received much less attention in the past. However, sound texture does not only play an important role in our perception of the environment but also used in various media compositions. In recent years, while the virtual reality and augmented reality technology become popular, it would be beneficial if the soundscape can be generated along with the visual scene. Unfortunately, it is not always easy to obtain the sound of a scene. Many of the virtual scenes are fictional. Recording sound effects can take a lot of effort, compared to that you can simply generate instrumental sounds with few clicks on laptop. The situation becomes worse if the sound needs to be interacted with visual events. For example, the engine sound of a racing car

should synchronize with the car motion on the visual scene. These problems have raised the demand for sound texture synthesis and transformation. However, these sounds are not simply noisy; each of them has various kinds of structures, like the quasi-periodic surge of the tidal waves and the whistling in the wind blowing. Besides, there are usually many random-occurring events inside, like the dripping in the rain or the crackling of the fire. These properties make sound textures very different from the instrumental sounds, and caused some problems when someone wants to manipulate sound textures with common existing algorithms. Therefore, a general mechanism, which designed to deal with sound textures, is vital. The goal of this research is thus to construct a framework which contains a model which describes the key characteristic of sound textures and allow the transformation of sound textures with the aid of the model. Furthermore, based on the model, one could seek the possibility to achieve morphing or interpolation between different sound textures. There are some potential application, although is not the aim of this research but still interesting, such as generate unseen sound textures for the artistic use.

1.3 Difficulties of Sound Texture Transformation

In present, there are many transformation algorithms which are capable of achieving high-quality transformations on sinusoidal sounds with perceptually strong sinusoidal components (Bloit et al., 2009, Dolson, 1986, Laroche and Dolson, 1999a,b, Serra, 1997), while also preserving the envelope and naturalness of transients (Röbel, 2003, 2010, Röbel and Rodet, 2005). While these algorithms can manipulate instrumental sounds properly, they yield less satisfying results when applied to sound textures or noises (Liao et al., 2012). For example, in (Liao et al., 2012), the traditional phase vocoder is not able to stretch a Gaussian noise properly.

The reason is that, the traditional phase vocoder heavily relies on the properties of sinusoidal sounds. When it comes to time stretching, it arouses two problems. First, since the phase vocoder assumes a sound is mainly composed by sinusoids, it will lengthen/shorten the existing peaks in the spectrogram and try to assign continuous phases to these components. This changes the structure of the sound texture and creating artifacts. Second, when humans perceive sounds with no significant harmonic structures and no sinusoidal components, the auditory system distinguishes them by other characteristics (McDermott and Simoncelli, 2011a). In the case of Gaussian noise, if the correlation between analysis frames is preserved after stretching, the result will be much more satisfying(Liao et al., 2012). Consequently, a different assumption about the sound and a different time stretching approach are required to transform sound textures

properly. There is a work which treats noises as a composition of short sinusoids (Hanna and Desainte-catherine, 2003). Another work(Apel, 2014) monitors sinusoidality when stretching a noise. These works provide another way to deal with the noise stretching problem. However, the structure of sound texture is more complicated than the structure of noise. Therefore, it is necessary to develop a comprehensive method to deal with sound textures.

1.4 Perception and Statistics

In this thesis, we would like to deal with the sound textures by preserving only the perceptual important properties during the transformation and synthesis. The reason is that, sound textures are generated by various physical processes. These processes can be very different from each other. Therefore, it is hard to find a general physical model within a reasonable complexity. Also, according to McDermott's research(McDermott and Simoncelli, 2011a), humans recognize different textures by a fixed set of perceptual properties. This makes the perceptual approach more appealing than the physically-informed approach when developing a general model for sound textures.

Following the observations in section 1.1, although sound textures are non-periodic, noisy and uncertain in detail, they still exhibit some characteristics over a moderate time period. In the case of visual textures, Julesz's conjecture (Julesz, 1962) and (Portilla and Simoncelli, 2000) suggests that these structures can be described by a series of statistical properties. On the other hand, (McDermott and Simoncelli, 2011a) and (McDermott et al., 2013), states that humans distinguish between sound textures based on time-averaged statistical properties in different auditory bands. It is also shown that, the longer the sound texture, the harder for humans to distinguish between different segments of the same texture(McDermott et al., 2013), which further justifies the result. Based on these results, it seems reasonable to seek a perceptual-based mechanism that achieves sound texture transformation and synthesis by preserving the important statistical properties. In the rest of the thesis, sometimes we'll use the term *sound texture sample* for convenience. It means a realization of a sound texture, or, in other words, a segment of the sound texture.

1.5 Signal Representation

According to (McDermott and Simoncelli, 2011a), humans are sensitive to the statistics in different auditory bands. It means that evaluating properties in the time-frequency

domain is a sensible option. This motivates to establish the statistics on a time-frequency representation(TFR). In this way, statistical properties can be measured and preserved in the sub-bands of a TFR without losing the perceptual relevance of the statistical description, but gaining on the other end the computational efficiency.

TFR is a time-frequency analysis tool, which provides a signal representation in the time-frequency domain, for example, the STFT spectrogram. It provides a view of signal simultaneously in both time and frequency. TFR is usually achieved by applying a time-frequency distribution to the signal(Cohen, 1995). Filter bank is one special form of the time-frequency distribution. In this thesis, we are particularly interested in those TFRs that can be implemented in the form of a filter bank. In the rest of the thesis, whenever we mention about TFR, we mean the filter bank based TFRs. For example, Short-Time-Fourier-Transform(Allen, 1977), Constant-Q Transform(Brown, 1991) and Gabor Transform(Feichtinger and Strohmer, 1998) can all be implemented in the form of a filter bank.

The idea is to evaluate statistics across the TFR and re-synthesise sounds from the representation. However, not every TFR can be a feasible choice for sound texture processing. A feasible TFR needs to fit several constraints. The constraints are described here briefly: First, invertibility, it must be invertible. It would be better if the TFR is re-synthesizable from only coefficient amplitudes. Since sound textures are generally not sinusoidal, assigning phases for these sub-band signals is not an easy task. As a result, it is important that, for the selected TFR, to have an efficient phase reconstruction algorithm. Second, data grid regularity, in order to evaluate correlations properly, the time interval between coefficients in a sub-band must be regular. Additionally, although not necessary, it is preferred that the time intervals keep the same across all sub-bands. This avoids the requirement of interpolation when evaluating the cross-correlation between sub-bands. The reason why these constraints are necessary is discussed in chapter 3.

At the first glance, the invertible CQT(Constant-Q Transform)(Holighaus et al., 2013), which could approximate the auditory bandwidth, seems to be a nice choice. CQT is a logarithmic frequency scale transform. The logarithmic scale fits the perception of auditory system better than the linear scale frequency. However, it seems lacking an efficient phase reconstruction algorithm. Besides, there is another problem related to the auditory bandwidth configuration: there are examples of sound textures that are severely perturbed when resynthesized using auditory bands or wavelet representation. An example is the sound texture produced by grains falling in cup; it produces an irregular narrow band noise with rather high central frequency. The high frequency resonance simply disappears from the resynthesized texture, for example, when using McDermott's algorithm.

Another option is the STFT(Short Time Fourier Transform)(Allen, 1977). All the sub-band coefficients of STFT have the same regular time interval; therefore, no interpolations are required. The linear frequency scale of STFT helps to detect the spectral dynamics of narrow band events in high frequency bands. Also, establishing a model based on the STFT representation can allow us to utilize existing efficient implementations of DFT(Discrete Fourier Transform). Due to these beneficial properties, the STFT will be used in the following investigation.

1.6 Perceptually Important Statistics of Sound Textures

Julesz's conjecture (Julesz, 1962) states that humans cannot distinguish between visual textures with identical statistics, the idea holds for most of the visual textures (Portilla and Simoncelli, 2000). This means that features of a visual texture could be well presented by a statistical model. It should be also true for sound textures if the perceptually important properties of sound textures are described by means of statistical features. Following this assumption, the perceptually important statistics could also characterize the randomized events and the structures which are created by those hidden stochastic processes in sound textures.

In the case of sound textures, statistics directly evaluated from the sound samples may not be perceptually meaningful. An example is that, randomly placed impulses with enough density would sound the same as a Gaussian noise. In this case, time domain moments are very different between these two sounds, but these sounds have similar moments in the time-frequency domain. This will be discussed in chapter 3 and depicted in Fig. 3.2. McDermott's work (McDermott and Simoncelli, 2011a) used a perceptual based ERB-filterbank(Glasberg and Moore, 1990) to divide a signal into several sub-bands. We call these divided signal components in each sub-band as **sub-band signals**. In the same article, he suggests that a proper description of a sound texture is composed of envelope statistics of sub-band signals. He also published an experiment result in (McDermott et al., 2013), further suggesting that time-averaged statistics are crucial to human perception of sound textures. From McDermott's works, he proposes that the perceptually important statistical properties consist of at least three components: moments, temporal correlation and spectral correlation.

In addition to the three perceptually important statistical properties, (Mesgarani et al., 2009) and (Depireux et al., 2001) suggest another property, the spectro-temporal correlation. These works explained why spectro-temporal ripples considered important to human when recognizing frequency modulated structures. In (Mesgarani et al., 2009), it states that the spectro-temporal stimuli reconstruction can be improved with the prior

knowledge of statistical regularities. Depireux's(Depireux et al., 2001) work indicates that a spectro-temporal envelope cannot always be separated into pure temporal and pure spectral response functions in the auditory cortex. This inseparability indicates that there exist some spectrally and temporally intertwined stages of processing in the auditory system. According to these experiment results, it seems reasonable to treat spectro-temporal correlations as one of perceptually important statistics.

The following subsections will briefly introduce these properties and discuss how we establish a statistical model based on these previous discoveries.

1.6.1 Moments

Moments, also known as marginal statistics(McDermott et al., 2009, Portilla and Simoncelli, 2000). According to McDermott's work(McDermott and Simoncelli, 2011a), the perceptually important moments are the moments of sub-band signal envelopes in the time-frequency domain. These moments describe information about the shape of histogram of sub-band signal envelopes(Lorenzi et al., 1999, Strickland and Viemeister, 1996). The first four moments are:

Mean $\mu = \frac{1}{N_x} \sum_{t=0}^{N_x-1} x(t)$

Variance $\sigma^2 = \frac{1}{N_x} \sum_{t=0}^{N_x-1} (x(t) - \mu)^2$

Skewness $\eta = \frac{1}{N_x \sigma^3} \sum_{t=0}^{N_x-1} (x(t) - \mu)^3$

Kurtosis $\kappa = \frac{1}{N_x \sigma^4} \sum_{t=0}^{N_x-1} (x(t) - \mu)^4$

The first moment(mean) describes the average, the second moment(variance) represents the spreading, the third moment describes the asymmetrical skew in the histogram and the fourth moment(kurtosis) describes the sparsity of the histogram(Attias and Schreiner, 1998, Field, 1987). Higher order moments such as hyper-skewness and hyper-flatness can describe more detailed information of the histogram. Moments are not always independent from each other, though skewness and kurtosis are independent from mean and variance, skewness and kurtosis do not independent from each other. An interesting statement is that the kurtosis is bounded below by the squared skewness plus 1(Pearson and Shohat, 1929). Though the usages of higher orders are possible,

their perceptual relevancy remain in question. McDermott tried to use all the moments in his model(McDermott and Simoncelli, 2011a), but the perceptual evaluation did not improve significantly. Portilla(Portilla and Simoncelli, 2000) also states that due to the extra complexity of imposing higher order moments, only the first four moments are used to form the marginal statistics in common cases.

1.6.2 Temporal Correlation

Temporal correlation refers to the correlation along time axis in the time-frequency domain, for example, the sound of helicopter rotor spinning. This can be interpreted as the autocorrelation function of a frequency band in a TFR. Sound textures are not periodic, but the appearance of recurring structure is still common. In Portilla's work(Portilla and Simoncelli, 2000) and an earlier work of McDermott(McDermott et al., 2009), autocorrelation was included. McDermott later dropped autocorrelation function due to the introduction modulation bands in his model(McDermott and Simoncelli, 2011a). The modulation bands characterize the envelope of the spectrum of a sub-band signal, this works similarly to the autocorrelation function(1.1), which is equal to the power spectrum of a signal.

Autocorrelation Function(ACF):

$$A_x(\tau) = \sum_{t=-\infty}^{\infty} x(t)\overline{x}(t+\tau) \tag{1.1}$$

1.6.3 Spectral Correlation

Spectral correlation is the cross-correlation between different frequency bands in the time-frequency domain. For example, the clicking of fire, which causes a sharp attack in all the frequency bands simultaneously. In contrast of the temporal correlation, which is a property of single frequency band, the spectral correlation represents the relation between a pair of frequency bands. Spectral correlation can be obtained from sub-band coefficients or features of sub-bands, like the envelope of cochlear bands (Glasberg and Moore, 1990). Cross-correlation(1.2) is an important factor in both Portilla(Portilla and Simoncelli, 2000) and McDermott's(McDermott and Simoncelli, 2011a, McDermott et al., 2009) works.

Cross-Correlation:

$$C_{x,y} = \sum_{t=-\infty}^{\infty} x(t)\overline{y}(t) \tag{1.2}$$

1.6.4 Spectro-Temporal Correlation

While the temporal correlation and spectral correlation characterize the horizontal and vertical relationships in the time-frequency representation, there are also spectro-temporal structures in the time-frequency representation. The slant relationship is the correlation involved in both time and frequency; one can consider it as a spectral correlation with a time delay. It is common that one frequency component is likely to appear after a brief delay of the appearance of another frequency component, for example, chirps and vibrating whispers.

In the time-frequency domain, spectro-temporal correlation can be characterized by the delayed cross-correlation, which are the terms with non-zero lags in cross-correlation functions(1.3).

Cross-Correlation Function(CCF):

$$C_{x,y}(\tau) = \sum_{t=-\infty}^{\infty} x(t)\overline{y}(t+\tau) \tag{1.3}$$

We know that humans mainly perceive envelopes and are sensitive to envelope statistics (Joris et al., 2004, McDermott and Simoncelli, 2011a). Therefore, we should focus on the statistics of coefficient magnitude in the time-frequency domain. In chapter 3, we will establish the statistical description based on these perceptually important properties proposed by (McDermott and Simoncelli, 2011a), (Mesgarani et al., 2009) and (Depireux et al., 2001).

1.7 Discussion

In this chapter, we have introduced the general idea of the sound texture:

> *A sound is a sound texture if human recognizes it through something else than time-limited characteristics and these characteristics must be the same for all segments of the same sound.*

The importance of sound texture in various applications raises the demand of sound texture processing algorithms. Sound textures have distinct nature from the instrumental sounds in both the signal structure and human perception. As a result, traditional algorithms cannot deal with the sound textures properly. In some previous works, people have found that sound textures behave like the visual textures in the sense that humans

recognize them through statistically related properties. The idea we suggest on sound texture processing is to characterize sound textures with a statistic-based description in the time-frequency domain.

The previous works(Depireux et al., 2001, McDermott and Simoncelli, 2011a, Mesgarani et al., 2009) have well explored the perceptual important statistical properties in the time-frequency domain. They suggest that the perceptually important statistics are the statistical moments and the spectro-temporal correlations in the time-frequency domain. Therefore, we do not seek another perceptual model for sound textures. Instead, we seek to propose a mechanism to analysis and synthesise sound textures efficiently. This will include a statistical description that characterizes the known perceptually important properties and an algorithm that efficiently resynthesize sound signals according to the statistical description. New samples of the sound texture can thus be generated from the statistical description. A time-frequency representation will be used as the underlying representation of the statistical description. Additionally, if certain criteria are met, any time-frequency representation can be used as the underlying signal representation. The statistical description serves as a generator that delivers different samples each time, but all these re-synthesized samples will perceptually belong to the same sound texture.

This work was funded by the French National Research Agency(ANR) under the PHYsically informed and Semantically controllable Interactive Sound synthesis(PHYSIS) project (ANR-12-CORD-0006) in the Contenus et Interactions (CONTINT) 2012 framework

Chapter 2

State of the Art

There are plenty of previous works on synthesizing sound textures. Some thorough reviews have been proposed(Schwarz, 2011, Strobl et al., 2006). Here, we roughly categorize them into model-based synthesis and granular-based synthesis. Model-based synthesis proposes various signal models for sound textures. Most of the models contain probabilistic factor. Some of the models are even physically-informed(Conan et al., 2013, Oksanen et al., 2013). The model serves as an interface between the analysis and synthesis of sound textures. On the other hand, granular-based synthesis treats sound textures as a superposition of basic atoms. The synthesis is achieved by reassembling, rearranging or reshuffling these atoms. The way we categorize sound textures is different from the way Schwarz(Schwarz, 2011) did. The reason is that, in this chapter we consider the difference between synthesizing from pure parametric signal model or synthesizing from atomic grains is the most important. Parametric models allow interpolation and possibly morphing of sound textures, which is also a future prospective of this research. In contrast, it is hard to interpolate/morph sound texture with granular synthesis algorithms. However, there exists some works that reside between these two categories. For example, Verron's(Verron et al., 2009) work is somewhat between these two. There are also outliers. For example, Schwarz's work(Schwarz, 2004) reassembles atoms from a corpus database instead of the input sample itself.

2.1 Early Attempts

Many of early attempts related to textures were aimed for visual textures. Many ideas about sound texture synthesis were burrowed from the domain of visual textures. As a result, these works are included here as the previous works.

One of the first complete investigations about how human recognize visual textures was proposed by Julesz(Julesz, 1962). In this work, he proposed Julesz's conjecture. It states that humans cannot distinguish between two textures that have the same second-order statistics. Although the conjecture was later proved false(Julesz et al., 1978), and in fact humans can distinguish between visual textures equated at all orders of statistics(Tyler, 2004). However, this idea was later burrowed by works in sound textures(McDermott and Simoncelli, 2011a).

One of the earliest model-based approaches to synthesis sound texture is proposed by Arnaud(Saint-Arnaud and Popat, 1998). In his work, he considers the sound texture as a composition of many basic elements, the atoms. This is a two-level representation, with the atoms being the first level and the probabilistic transitions between atoms being the second level. The atoms are the audible sound events in a sound texture, like the individual raindrops in a rain texture. A texture may contain several different kinds of atom. After identified the atoms from a sample of sound texture, the probability distribution or the probabilistic transition between atoms are estimated. The synthesis process then imitates the probabilistic distributions and transitions of atoms to generate a new instance of the input sound texture. The algorithm is straightforward, but there are difficulties when someone intends to use the algorithm to generate sound textures. First, it is not easy to correctly identify the occurrence of an atom and separate it from the rest of the signal. Another difficult task is how to categorize these individual atoms. Many atoms may look similar, but never the same. Despite of the difficulties mentioned above, Arnaud's work is one of the first attempts which used statistical concepts on sound texture synthesis. The two-level representation also influenced many other works.

There's an important parametric model for image textures proposed by Portilla(Portilla and Simoncelli, 2000). He proposed a texture model based on the statistics of wavelet coefficients. The model was inspired by Julesz's conjecture and works on joint statistics. The statistics he used included autocorrelation, cross-correlation and moments. The statistics are imposed by a gradient projection algorithm. His work influenced McDermott's thought on sound textures.

2.2 Model-Based Synthesis

Most of the works about sound texture synthesis fall into this category. However, the variety between these models is large. In (Hanna and Desainte-catherine, 2003), stochastic noises are synthesized with randomized short sinusoids. The intensity of noise is kept during the transformation by preserving the statistical moments with the short sinusoids. (Athineos and Ellis, 2003) uses linear prediction in both time and frequency

domain(TFLPC), preserving the short-term temporal correlations in order to synthesize the brief transients in sound texture. Zhu(Zhu and Wyse, 2004) proposed another TFLPC-based approach with a two-level foreground-background representation. The foreground is the transients and the background is the remainder. The two levels are synthesized separately then mixed together. The two-level approach also appears in the environmental sound scene processor proposed by (Misra et al., 2006). Verron(Verron et al., 2009) proposed a parametric synthesizer which was based on additive synthesis. If proper parameters are given, the synthesizer is capable of generate environmental sounds such as rain or fire from limited number of basic sounds. Bruna(Bruna and Mallat, 2011) proposed a new wavelet transform which provides a better view of textures while capturing high-order statistics. Kersten(Kersten and Purwins, 2012)'s work aimed to re-synthesis the fire crackling sound with a model similar to the common foreground-background model(Athineos and Ellis, 2003, Misra et al., 2006, Zhu and Wyse, 2004).

McDermott(McDermott et al., 2009) proposed a model which adapted from (Portilla and Simoncelli, 2000), which resynthesizes target sound textures with high order statistics and spectral correlations between sub-band signal envelopes. In his article, statistical properties are applied to Gaussian noises to generate different sound textures. The study of McDermott was the first to support the view that sound texture perception is related to the moments and correlations that may be calculated in different locations of the auditory system. Later in (McDermott and Simoncelli, 2011a), the model was further refined. The model uses ERB-filterbank to divide the signal into perceptual bands. Within each perceptual band, each band was further divided into smaller modulation bands. The statistics he uses includes the first four moments of each modulation band and perceptual band, along with the cross-correlations between neighbouring perceptual bands and modulation bands.

There are some other works seek to develop a high quality synthesize algorithm for a specific sound texture by studying the physical process which generates the sound. These approaches analyse the physical processes which induced the sound texture, then use the result as a footstep to develop algorithms. In some way, it is possible to analyse the physical process of a specific kind of sound texture then develop high quality synthesis and transformation. (Oksanen et al., 2013) used parallel waveguide models to synthesis the sound of jackhammers. His work is capable of synthesizing a series of jackhammer impacts. (Conan et al., 2013) proposes a model that characterizes rubbing, scratching and rolling sounds as series of impacts. The work comes with a schematic control strategy to achieve sound morphing between these sounds by preserving the invariants between different kinds of sounds.

2.3 Granular-Based Synthesis

Rather than proposing a model or dealing with the statistics directly, some works aim to resynthesis sound textures with granular synthesis. Dubnov (Dubnov et al., 2002) processes a sound texture in the form of wavelet tree and reshuffles sub-trees if their difference is smaller than a threshold. This results into a new sample of the texture, which the new sample is fully composed by the rearranged segments in the original sample. Fröjd(Fröjd and Horner, 2009) proposed a straightforward approach. Several blocks are cut from the original sound texture. New samples are generated by rearranging and cross-fading these blocks. O'Leary(O'Leary and Robel, 2014) took a path similar to Dubnov to synthesis sound textures without using wavelets. In his work, he searches atoms in the time-frequency domain by evaluating the correlations. The algorithm then finds a proper point to cut and rearrange these atoms. More variety can be achieved if atoms can also be replicated instead of only shuffling. He called this mechanism the 'montage approach'. Both of these algorithms create new samples with very little to none artifacts. This is a great advantage for this kind of algorithms.

On the other hand, Schwarz(Schwarz, 2004) proposed a different approach. He proposed a descriptor-driven, corpus-based approach to synthesize sound textures. The input texture was first transformed into audio descriptors, then the synthesis proceeds by combining sounds selected from the corpus database. The sounds are selected such that the combination of these sounds fit the audio descriptors. His work is more close to an orchestration system dedicated for sound textures. Later he proposes a statistical model(Schwarz and Schnell, 2010) which uses histogram and Gaussian mixture model to model the descriptors. This model enhances the controllability of his corpus-based synthesis.

2.4 Discussion

One perspective of this research is to achieve sound texture transformation such as sound texture morphing. Under the circumstance, model-based approach is the most promising choice. Granular-based synthesis is powerful in delivering high quality samples with little no none artifacts, but the parametric nature of model-based synthesis has much more flexibility. Moreover, if the signal model is powerful enough, model-based approach can also deliver high quality results with no artifacts. Among these model-based approaches, we are specifically influenced by McDermott's work(McDermott and Simoncelli, 2011a). Although the work was originally modified from a successful visual texture algorithm, McDermott put a lot of effort in adapting the model to fit human auditory system. The

perceptual important statistics in his model have plentiful support from publications in the related research field.

The work of McDermott discovers a comprehensive set of perceptual important statistics and delivers high quality synthesized results. However, his work did not aim to create an efficient sound synthesis algorithm, but aim to study the perception of texture statistics. In an experimental investigation, we have found that the algorithm(thanks to him for the original code) is limited by some approximations. For example, to keep the sub-band signals band-limited, he applied sine filters to them after each iteration, which introduces artifacts. The algorithm also takes a while to generate new sound texture samples. Therefore, we would like to propose a mechanism that is capable of producing new sound texture samples efficiently but the perceptual quality of the new samples still rival to McDermott's results.

Chapter 3

Statistical Description of Sound Textures over TFR

In the previous section (sect 1.1), we know that there's no widely accepted definition of sound textures. However, we had the idea that: *A sound is a sound texture if human recognize it through something else than time-limited characteristics and these characteristics must be the same for all segments of the same sound.* This idea serves as the fundamental assumption for the whole thesis. Fortunately, some previous works (McDermott and Simoncelli, 2011a, Mesgarani et al., 2009) already shown a strong evidence about what are the perceptually important characteristics to the human auditory system. These characteristics are: spectral correlation, temporal correlation, spectro-temporal correlation and moments. Therefore, we would like to propose a statistical description based on these perceptually important properties proposed by (McDermott and Simoncelli, 2011a), (Mesgarani et al., 2009) and (Depireux et al., 2001).

Combining the idea with the fact that human perception of sound texture is heavily related with perceptual important statistics such as auto-correlation, cross-correlation and moments, the idea can be described in a more definitive way:

Definition (Sound Textures). *If a sound is a sound texture, any segment of this sound has the property that its expectation values of auto-correlation function, cross-correlation function and moments in time-frequency domain do not dependent on the temporal position of the segment.*

Even though the definition above can be applied to most of the sound textures, not all the 'sound textures' fit into this definition, for example, the sound of tidal waves and people babbling in a cocktail party. These sounds are generally classified as sound textures, but human can clearly distinguish the events in the sounds. In our opinion,

if human can identify a specific event in a sound, the sound is more likely to be a mixture of events and textures. Nonetheless, this article would focus on those sounds, which fit into the definition. Since those sound textures which fit the definition can all be described with time-invariant properties(Julesz, 1962, McDermott and Simoncelli, 2011a). Therefore, a statistical description defined on the time-frequency domain should be a legitimate model to serve as the basis of analysis-synthesis scheme of sound textures. In this chapter, we start with the selection of a viable filter bank based TFR, and then followed by an overview of the statistical description. In the end, a time-frequency domain statistical description is proposed, along with the steps about how to obtain the description from an arbitrary sound texture.

3.1 Selection of Time-Frequency Representation

Time-Frequency Representation(TFR) is a two-dimensional signal representation over both time and frequency(Cohen, 1995). It provides a way to analyze signals in the time-frequency domain. In this thesis, we discuss only a subset of TFRs that are achievable by filter banks. TFRs are usually composed of complex-valued coefficients over the time-frequency domain. The columns are also known as analysis frames.

The selection of time-frequency representation for sound texture processing is not a trivial issue. In the case of processing speech or instrumental sounds, STFT is a suitable choice. The linear frequency scale of the STFT is convenient when dealing with the speech/instrumental harmonic structures. However, since most of the sound textures do not have harmonic structures, the feasibility of STFT should be re-considered. Another problem is that, from McDermott's work(McDermott and Simoncelli, 2011a), his experiments were conducted with perceptual bands, which is a logarithmic frequency scale. It is important to know whether his conclusions also applicable with linear frequency scales.

The next subsections will discuss what kind of TFRs are suitable to be the basis of a time-frequency domain statistical description for sound textures.

3.1.1 Invertibility

Invertibility means that signals can be resynthesized from the coefficients of the TFR. However, not all TFRs are invertible. For example, the original Constant-Q Transform(CQT)(Brown, 1991) is not invertible. Moreover, what we know is that human auditory system can sense the difference of amplitude statistics in the time-frequency

domain(Joris et al., 2004, McDermott and Simoncelli, 2011a). It means that preserving the magnitude of TFR coefficients is sufficient. However, if the TFR is complex-valued, we'll have to know how to assign the phase of TFR coefficients properly. As a result, we will prefer to choose a TFR that has an efficient algorithm to reconstruct the signal with only the magnitudes of TFR coefficients. In this regard, STFT is advantageous. As long as the overlap between neighbouring analysis frames is bigger than 50%, the information contained by the STFT magnitude is sufficient to represent a signal(Nawab et al., 1983). The resynthesis can be done by assigning phase values to the amplitude coefficients such that phase inconsistencies between neighbouring analysis frames are minimized. There exists many efficient algorithms which can assign coherent phase values from only the magnitude spectrograms (Griffin and Lim, 1984, Le Roux et al., 2008, Nawab et al., 1983, Zhu et al., 2006). On the other hand, the invertible CQT(Dörfler et al., 2003), to the best of our knowledge, has no efficient resynthesise algorithm from only magnitude coefficients.

3.1.2 Data Grid Regularity

Most of the traditional TFRs have a regular sampling interval across the same frequency bin. The time difference between two horizontal neighbouring coefficients is always the same, like STFT and CQT. Some TFRs do not follow this rule, for example, the non-stationary Gabor frame(Balazs et al., 2011) and the adaptive spectrogram(Chi et al., 2005, Liuni et al., 2011). In order to evaluate autocorrelation and cross-correlation functions, we will need the regularity applies on both time and frequency axis. That is, a regular sampling interval, which applies to all frequency, bins, like the STFT. The condition may sound strict, but many irregular TFRs can achieve this with configuration changes at the cost of extra computation including most of the Wavelet transforms. For example, the invertible CQT combined with the non-stationary Gabor frame (ERBlet constant-Q transform via non-stationary Gabor filterbanks)(Necciari et al., 2013).

3.1.3 STFT v.s. invertible ERBlet CQT

It seems that STFT fits both criteria and is less computational intensive. With the conditions above, STFT seems to be a reasonable choice. The only problem lies in the linear frequency scale of STFT. All the McDermott's theories are based on perceptual frequency band, which is a logarithmic frequency scale. However, we can show that, if a perceptual band is divided into several smaller sub-bands, preserving the correlation of the sub-bands will also preserve the autocorrelation of the perceptual band.

If a real signal $s = x + y$, preserving the correlation functions of x and y will also preserve the auto-correlation of $s(x, y$ are real). Consider the equation below:

$$\begin{aligned} A_{x+y}(\tau) &= \int_{-\infty}^{\infty} (x(t) + y(t))(\overline{x}(t+\tau) + \overline{y}(t+\tau))\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} (x(t)\overline{x}(t+\tau) + y(t)\overline{x}(t+\tau) + x(t)\overline{y}(t+\tau) + y(t)\overline{y}(t+\tau))\,\mathrm{d}t \quad (3.1) \\ &= A_x(\tau) + C_{y,x}(\tau) + \overline{C_{x,y}}(\tau) + A_y(\tau)\ ,\ x \in \mathbb{R}, y \in \mathbb{R} \\ &= A_x(\tau) + A_y(\tau) + 2C_{x,y}(\tau) \end{aligned}$$

According to the result of (3.1), preserving the autocorrelation and cross-correlation of individual elements will also preserve the autocorrelation function of the summation. It means that, if a perceptual band is divided into several linear frequency bands, preserving the correlation functions of these bands will also preserve the autocorrelation function of the perceptual band. This result strengthens the feasibility of linear frequency scales. However, if the frequency resolution is too low in a linear frequency setup, one linear frequency band may contain multiple perceptual bands in low frequency parts. In this case, preserving the correlation of the linear frequency band cannot guarantee the correlation functions for those underlying perceptual bands. Therefore, the frequency resolution should be chosen such that the bandwidth of each bin is not greater than the narrowest perceptual band.

Another good choice is the invertible ERBlet CQT(Necciari et al., 2013). It has a logarithmic scale frequency, which fits the auditory perception. A setup which satisfies the data grid regularity can be done with the aid of LTFAT toolbox(Průša et al., 2014). The resynthesized ERBlet CQT spectrogram is shown in Fig. 3.1. Unfortunately, even though the proposed algorithm does properly generated magnitudes for the TFR, we cannot assign proper phase values for it. Conventional phase reconstruction algorithms do not work for the invertible ERBlet CQT. Therefore, in the end, we select STFT as the base TFR for the statistical description to achieve sound texture analysis/synthesis. However, readers should notice that, the proposed statistical description can be applied on any TFR which satisfies the two criteria above.

3.2 Overview of the Statistical Description

In this section, we would like to discuss the statistical description which can model most of the sound textures, or at least those sound textures which fits in our definition at the beginning of chapter 3. In section 1.6, it describes the perceptually important

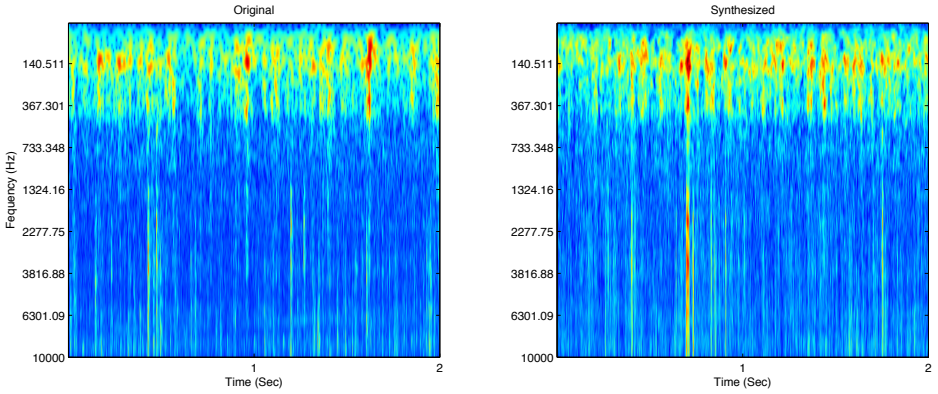


FIGURE 3.1: *The ERBlet CQT spectrogram of the fire texture. Left is the spectrogram of the original fire. Right is the spectrogram generated by the proposed algorithm*

statistics. It also suggests that these time-averaged statistics are sufficient to capture all those critical features that human uses to recognize different sound textures. While the correlation functions characterize the dependencies of all directions over the TFR plane, the moments help to characterize the shape of histogram of each frequency band.

Besides the statistics in the time-frequency domain proposed by (McDermott and Simoncelli, 2011a), there are also some works that employed statistics in the time domain. For example, (Hanna and Desainte-catherine, 2003) used statistical moments in the time domain. Time domain moments might be a useful predictor to ensure the quality of resynthesized noise, but it may not be the case when it comes to the human perception of sound texture. Consider a Gaussian noise and a signal, which is composed of randomized square pulses. These two sounds are perceptually similar to a certain degree. These two sounds have very different histograms in time domain. The sparsity differs in great scale, so does the kurtosis (Fig. 3.2). However, their band-wise histograms in the time-frequency domain are not that different. The band-wise kurtoses of coefficient amplitudes are much closer. It seems that the time-frequency domain kurtosis fits the human perception better. McDermott's (McDermott and Simoncelli, 2011a) work also mentioned that humans discriminate sound textures by sensing the change of envelope statistics in different frequency bands in the time-frequency domain. The statistics of the signal waveform is not perceptually important. As a result, no time domain statistics will be included.

In the preliminary stage of the research, the complex-valued cross-correlation functions of raw TFR coefficients were included in the statistical description (Liao et al., 2013). It means that the structures of phase values were put into account. At the time, we used Griffin&Lim's algorithm(Griffin and Lim, 1984) to achieve phase coherencies between neighbouring analysis frames. If the initial phase values were assigned properly, the result of Griffin&Lim's algorithm will significantly better. In the later stage of the

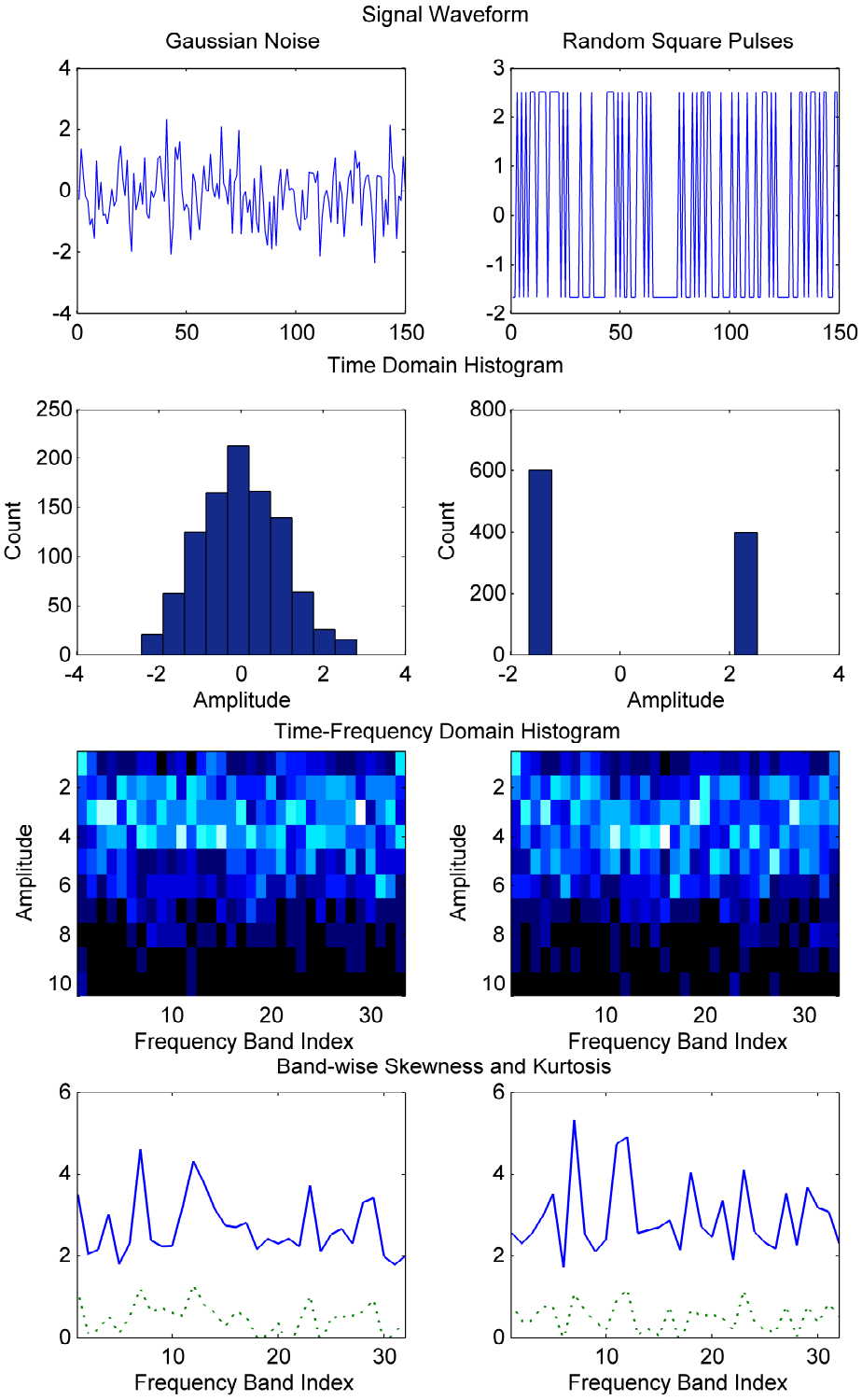


FIGURE 3.2: *The comparison between time-domain and time-frequency domain histogram. First row shows the first 150 samples of two 1000-sample signals(left:Gaussian, right:Square pulses). Second row is the histogram of waveform amplitudes(left:* $\eta = 0.14, \kappa = 2.73$*, right:* $\eta = 0.41, \kappa = 1.17$*). The third row is the band-wise histogram in the time-frequency domain, a brighter color indicates a higher count, and darker blue indicates a lower count. The fourth row plots band-wise skewness(green dotted) and kurtosis(blue solid).(left:* $\mu_\eta = 0.5, \mu_\kappa = 2.72$*, right:* $\mu_\eta = 0.52, \mu_\kappa = 2.95$*).*

research, phase coherency is achieved with LeRoux's algorithm(Le Roux et al., 2010), which is less reliant on the initial phase values. The advantages of using LeRoux's algorithm will be discussed in a later section.

In 3.1.3, we chose STFT as the underlying TFR. Now we are going to establish the statistical description on it. Each frequency bin of STFT is considered as a single sub-band. The coefficient magnitudes of each sub-band are the sub-band signals. Two sets of statistics are used to characterize each sub-band signal. The histogram of each sub-band signal is characterized by its first four central moments, namely mean, variance, skewness and kurtosis. The temporal correlation of the sub-band signal is characterized by the autocorrelation function. The relations between sub-band signals are characterized by cross-correlation functions.

There is something to be mentioned when transforming signals into TFR. The length and the hop size of analysis window must be chosen wisely. While the length controls the trade-off between time and frequency resolution, the hop size controls the sampling interval of correlations on the time-frequency domain. In one hand, if the hop size was too small, the computation required to evaluate the TFR and the statistics will be intensive. On the other hand, if the hop size was too large, the sampling interval of correlation functions will be large. If the sampling interval is too large, the shape of correlation function will be rough, possibly missing most of the detail parts. The quality of resynthesized sounds will thus degenerate. In our experiment, we found a window length of $128ms$ and a hop size of $8ms$ is appropriate in most of times when using STFT as the underlying TFR.

In order to achieve a proper resynthesis, the coherency between analysis frames must be preserved, while it can be more or less characterized by the time-averaged statistics like temporal correlations, it cannot be perfectly reconstructed. This should be mitigated in the phase reconstruction stage. Rather than the traditional Griffin& Lim's (Griffin and Lim, 1984) algorithm, LeRoux's algorithm(Le Roux et al., 2008) (Le Roux et al., 2010) is used to reconstruct the phase. Besides the concern of efficiency, the phase reconstruction of LeRoux prioritizes the phase consistency between neighbouring large coefficients, thus enforcing the inconsistencies settled between the smaller coefficients. This is advantageous due to sound textures often has noise components and inconsistencies hidden in these small components are hard to perceive.

Here is a brief summary of the statistics we use: the auto-correlation functions of sub-band signals, the cross-correlation functions of sub-band pairs and the first four sub-band moments. It can be written in the form of (3.2).

$$\Phi \equiv \{\mathcal{A}, \mathcal{C}, \mathcal{M}\} \tag{3.2}$$

3.3 Evaluate Statistics from TFR

In this section, we describe how the statistical description can be. We will begin from describing the mechanism that is used to extract statistics from sound textures. Assuming a sound texture x is given as input, we first apply a STFT with Fourier transform size N_k on x and get its STFT representation $X(n,k)$.

$$X(n,k) = \sum_{t=-\infty}^{\infty} x(t)w(t-nh)e^{\frac{-j2\pi tk}{N_k}} \tag{3.3}$$

Where n is the center position of the analysis frame in time domain, along with a window function w and hop-size h. We use Hanning window here since it has small side-lobes. The only trade-off is a slightly wider main-lobe(Harris, 1978). This helps us reducing the contamination in the cross-correlation function caused by the aliasing effect. In the digital application where the time is discrete, in order to cope with the assumption of the phase reconstruction algorithm(Le Roux et al., 2008), the window function must be made into the so called 'periodic' window.

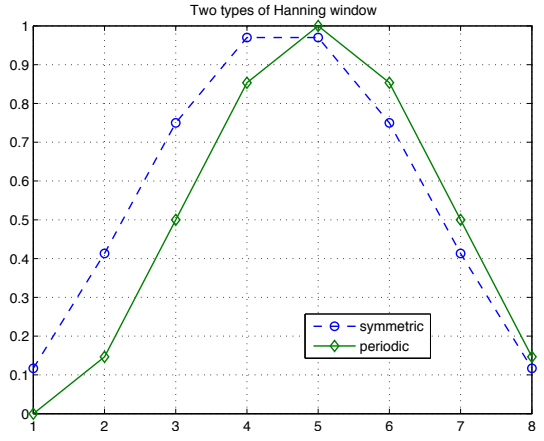


FIGURE 3.3: *The comparison between symmetric(blue dashed) and periodic(green solid) window*

Example: Periodic Hanning Window

$$w(n) = \frac{1}{2}(1 - \cos(2\pi\frac{n}{N})) \ , \ 0 \leq n \leq N \ , \ N \, even \tag{3.4}$$

This can also be achieved by deleting the right-most coefficient of an odd-length, symmetrical window, shown in fig 3.3. Empirically, a suitable windows function has a length between $64ms$ to $128ms$. Besides the window function, the hop-size also has to be chosen properly. Hop-size determines how many new samples are fed in the next analysis frame. The value of hop-size controls the rate of down sampling of all the correlation

functions. Higher rate of down sampling not only reduces the amount of computation, but also removes the fine detail from the correlation functions. Empirically, a hop-size with effective time-scale between $8ms$ to $16ms$ is an appropriate choice. Next, we keep only the magnitude of the STFT and apply an amplitude compression(McDermott and Simoncelli, 2011a, Ruggero, 1992). Then we have the sub-band signal $\mathcal{X}_k(n)$:

$$\mathcal{X}_k(n) = |X(n,k)|^{2c} \tag{3.5}$$

In (3.5), we applied a stronger amount of compression($c = 0.15$) than the amount in the auditory system($c = 0.3$)(Moore, 2007). In the practical situation, we found that $c = 0.15$ helps to avoid some numerical problems and improves the quality of the result. The numerical problem will be described later. Then, the perceptual important statistics are evaluated on $\mathcal{X}_k(n)$. To be mentioned, except for the mean, we store central moments in the statistical description. The central moments are equal to the standardized moments without denominators. The central moments become equivalent to the standardized moments when the variance of the signal is 1.

Moments

$$\mathcal{M}_{\mathcal{X}} = \left\{ \begin{pmatrix} \mu(\mathcal{X}_0) \\ \mu_2(\mathcal{X}_0) \\ \mu_3(\mathcal{X}_0) \\ \mu_4(\mathcal{X}_0) \end{pmatrix}, \begin{pmatrix} \mu(\mathcal{X}_1) \\ \mu_2(\mathcal{X}_1) \\ \mu_3(\mathcal{X}_1) \\ \mu_4(\mathcal{X}_1) \end{pmatrix}, \cdots, \begin{pmatrix} \mu(\mathcal{X}_K) \\ \mu_2(\mathcal{X}_K) \\ \mu_3(\mathcal{X}_K) \\ \mu_4(\mathcal{X}_K) \end{pmatrix} \right\} \tag{3.6}$$

Autocorrelation Functions

$$\mathcal{A}_{\mathcal{X}} = \{A_{\mathcal{X}_0}, A_{\mathcal{X}_1}, \cdots, A_{\mathcal{X}_K}\} \tag{3.7}$$

Cross-correlation Functions

$$\mathcal{C}_{\mathcal{X},\mathcal{X}} = \begin{Bmatrix} C_{\mathcal{X}_0,\mathcal{X}_1} & C_{\mathcal{X}_0,\mathcal{X}_2} & \cdots & C_{\mathcal{X}_0,\mathcal{X}_K} \\ & C_{\mathcal{X}_1,\mathcal{X}_2} & \cdots & C_{\mathcal{X}_1,\mathcal{X}_K} \\ & & \ddots & \vdots \\ & & & C_{\mathcal{X}^{K-1},\mathcal{X}_K} \end{Bmatrix}, \; K = \left\lfloor \frac{N_k}{2} \right\rfloor \tag{3.8}$$

Finally, the extracted statistical description $\Phi(\mathcal{X})$ can be written as:

$$\Phi(\mathcal{X}) \equiv \{\mathcal{M}_{\mathcal{X}}, \mathcal{A}_{\mathcal{X}}, \mathcal{C}_{\mathcal{X},\mathcal{X}}\} \tag{3.9}$$

In (3.9), $\mathcal{A}_{\mathcal{X}}$ denotes the autocorrelation functions of $\mathcal{X}$, $\mathcal{C}_{\mathcal{X},\mathcal{X}}$ denotes the cross-correlation functions and $\mathcal{M}_{\mathcal{X}}$ denotes the first four central moments. The representation is in fact over-determined, therefore the statistical information captured by cross-correlation functions are redundant. In order to reduce the amount of parameters and decrease the computational intensity, only a subset of cross-correlation functions is selected. In theory, the amount of cross-correlation functions can be reduced to such that $\mathcal{C}$ include only the cross-correlation functions of neighbouring sub-band signals. In the practical case, a cross-correlation function is selected if:

a) The two sub-bands are adjacent to each other.

OR

b) The index difference between the two sub-bands is a multiple of 3.

3.4 Discussion

In this chapter, we have discussed about the choice of TFR. STFT is not the best choice in terms of the correspondence to the perceptual frequency bank. However, it requires only little computation power compared to other more complicated TFRs such as ERBlet CQT(Necciari et al., 2013). The most important thing is that the phase reconstruction and coherency problem on STFT has been studied for a long time, thus having plenty of efficient algorithms. If an efficient phase reconstruction algorithm can be found on the ERBlet CQT, it would be a competitive choice as a base TFR.

In the second half of the chapter, we established a statistical description on the STFT. The statistical description characterizes envelope statistics on the time-frequency plane. The basic idea of resynthesis would be generate a signal which fits the statistics in the description. The next chapter will give some background knowledge regarding to imposing the statistics to an arbitrary signal.

Chapter 4

Imposing Statistics

This chapter introduces several ways to impose statistical properties to a given TFR. The impositions of three types of statistical properties are described: autocorrelation functions, cross-correlation functions and central moments. In some circumstances, one can impose statistics without iterative optimization process. However, if high quality synthetic samples are desired, iterative optimization is required for most of the time. We will also talk about the difficulty encountered when dealing with complex objective functions. A problem related to these imposition methods is that the sub-band signals are originally non-negative, but none of these algorithms ensures non-negativity of the result. Fortunately the problem is mitigated with the magnitude compression in (3.5). In the end, some of the imposition algorithms are selected to be a part of the analysis-synthesis framework. The complete steps about how to generate a new sample of texture will be described in the next chapter.

4.1 Full Imposition of Correlation Functions

States-of-the-art algorithms use iterative optimization process to impose autocorrelation and cross-correlation functions(McDermott and Simoncelli, 2011a, McDermott et al., 2009, Portilla and Simoncelli, 2000). This is not efficient since that whenever a TFR coefficient is changed, all the autocorrelation and cross-correlation functions need to be re-evaluated. The re-evaluation is quite computation intensive. It would be beneficial if we find another formulation of the optimization problem and reduce the amount of re-evaluation in each iteration.

It is widely known that the evaluation of correlation functions can be achieved by Fourier transform. This is based on the Wiener-Khinchin theorem and the cross-correlation theorem:

$$\hat{A}_{\mathcal{X}} = \mathcal{F}(\mathcal{X}(t) * \overline{\mathcal{X}}(-t)) = |\hat{\mathcal{X}}|^2 \tag{4.1}$$

$$\hat{C}_{\mathcal{X},\mathcal{Q}} = |\hat{\mathcal{X}}||\hat{\mathcal{Q}}|e^{j(\theta(\hat{\mathcal{X}})-\theta(\hat{\mathcal{Q}}))} \tag{4.2}$$

Here, both the hat symbol $\hat{}$ and $\mathcal{F}$ mean the Fourier transform, and θ means the principal angle of each element in the complex vector. Equation (4.1) shows the duality between autocorrelation function and power spectrum. However, using the spectrum of correlation function calculates a circular cross-correlation function on the signal. This induces the circular aliasing effect. The circular aliasing effect takes place whenever a signal x is not time limited to less than half of the segment length. In fact, the circular aliasing effect is equivalent to the superposition of the autocorrelation function of the zero-padded version of signal x. Suppose x_z is the zero-padded version of signal x and the length of x_z is twice as long as the length of x, we can depict the circular aliasing effect by the following equation:

$$\begin{aligned} A_x(\tau) &= A_{x_z}(\tau) + A_{x_z}(\tau - N_x) \\ &= A_{x_z}(\tau) + A_{x_z}(N_x - \tau), \quad 0 \le \tau < N_x \end{aligned} \tag{4.3}$$

In (4.3), if $A_{x_z}(\tau) \gg A_{x_z}(N_x - \tau)$, then $A_x(\tau)$ is not strongly affected by the circular aliasing. According to the sound texture samples we have, the power of $A_{x_z}(\tau)$ is at least ten times larger than the power of $A_{x_z}(N_x - \tau)$ in about half of the autocorrelation terms. These terms constitute about 80% of the energy of autocorrelation functions. Therefore, the circular aliasing effect does not pose a strong effect in general.

Next, recall from section 1.6, sub-band signals are the magnitude coefficients in a frequency bin of TFR. Consider that we have two sub-band signals $\mathcal{X}, \mathcal{Q}$, then combine (4.1) with (4.2), we have:

$$\sqrt{\hat{A}_{\mathcal{X}}} = |\hat{\mathcal{X}}| \tag{4.4}$$

$$\theta\left(\frac{\hat{C}_{\mathcal{X},\mathcal{Q}}}{\sqrt{\hat{A}_{\mathcal{X}}}\sqrt{\hat{A}_{\mathcal{Q}}}}\right) = \theta(\hat{\mathcal{X}}) - \theta(\hat{\mathcal{Q}}) \tag{4.5}$$

From (4.4) and (4.5), it can be seen that, when the autocorrelation function is fixed by the spectral magnitude of the signal, cross-correlation function can be controlled by changing the size of included angles between two spectra. The included angles are

simply the phase differences between $\hat{\mathcal{X}}$ and $\hat{\mathcal{Q}}$. In other words, changing the cross-correlation function by alternating the spectral phase differences between two signals will still preserve the autocorrelation function. Now, we can represent a sub-band signal in the form of its correlation functions:

$$\mathcal{X}_k = \mathcal{F}^{-1}\left\{\sqrt{\hat{A}_{\mathcal{X}_k}}e^{j\left(\theta(\frac{\hat{C}_{\mathcal{X}_k,\mathcal{X}_{k-1}}}{\sqrt{\hat{A}_{\mathcal{X}_k}}\sqrt{\hat{A}_{\mathcal{X}_{k-1}}}})+\theta(\hat{\mathcal{X}}_{k-1})\right)}\right\} \tag{4.6}$$

If $\mathcal{X}_0$ is real, then $\theta(\hat{\mathcal{X}}_0)$ has to be Hermitian($f(t) = \overline{f}(-t)$) and that the value for the DC-term(0th term) and the Nyquist frequency term must be real. In (4.6), if we determined the phase of $\hat{\mathcal{X}}_0$, or $\theta(\hat{\mathcal{X}}_0)$, we can continue to determine all the sub-band signals. According to (4.6), the imposition of correlation functions thus can be done by first replacing the magnitude spectra of all sub-band signals by the magnitude spectra of the input sound texture. Then, set the phase differences of the neighbouring spectra respect to the phase differences of the corresponding neighbouring spectra of the input sound texture.

(4.6) can also be understood in another way. It can be considered that (4.6) is a process of adding phase offsets to the frequency bins of $\hat{\mathcal{X}}_k$, which is the spectra of the original sub-band signal $\mathcal{X}$. It is an offset that applies to the same bin of all the sub-band signals, thus not changing any phase differences. Each different bin receives a different offset, thus the output of (4.6) is totally determined by a *phase offset vector*, θ_{offset}, with length $\lfloor N_k/2 \rfloor + 1$. In this way, the procedure looks like generating a new sound texture sample, $\tilde{\mathcal{X}}$, from the original one, as shown in (4.7).

$$\begin{aligned}\tilde{\mathcal{X}}_k &= \mathcal{F}^{-1}\left\{\sqrt{\hat{\mathcal{A}}_{\mathcal{X}_k}}e^{j\left(\theta(\frac{\hat{C}_{\mathcal{X}_k,\mathcal{X}_{k-1}}}{\sqrt{\hat{A}_{\mathcal{X}_k}}\sqrt{\hat{A}_{\mathcal{X}_{k-1}}}})+\theta(\tilde{\hat{\mathcal{X}}}_{k-1})\right)}\right\} \\ \tilde{\mathcal{X}}_0 &= \mathcal{F}^{-1}\left\{\sqrt{\hat{\mathcal{A}}_{\mathcal{X}_k}}e^{j\left(\theta(\hat{\mathcal{X}}_0)+\theta_{offset}\right)}\right\}\end{aligned} \tag{4.7}$$

Equation (4.6) established a direct connection between the signal representation and the statistical properties (correlation functions). Moreover, according to the Parseval's theorem, preserving the autocorrelation will also preserve mean and variance of the signal, which means all second-order statistics are preserved. In fact, (4.6) is also a representation of a set of feasible sub-band signals. Every element in this set will have

statistics matching all the correlations and second order moments. (4.6) was employed in an earlier stage algorithm of this research(Liao et al., 2013).

As we mentioned in section 3.3, while we have only $\lfloor N_k/2 \rfloor + 1$ distinct sub-band signals, we have $\lfloor N_k/2 \rfloor (\lfloor N_k/2 \rfloor + 1)/2$ different cross-correlation functions, thus the system is over-determined. Since it is an over-determined system, arbitrarily created cross-correlation functions will easily become inconsistent. One could only use the formula to generate signals with the designated correlation functions, under the condition that he could correctly define a set of consistent correlation functions. In our case, since the statistics are obtained from a real sound texture, we know there will be no inconsistencies between these cross-correlation functions. In theory, the minimum number of cross-correlation function we need is at least $\lfloor N_k/2 \rfloor$. This is the smallest amount required to generate the complete TFR with (4.6). There's another thing need to be concerned while selecting cross-correlation functions, if the spectrum of a cross-correlation function contains zero, one will not be able to find an unique solution with (4.5). In practice, the way we choose cross-correlation functions is according to the manner described in section 3.3. It selects the cross-correlation functions of neighbouring sub-band signals and those with index differences are the multiples of 3(e.g. $C_{\mathcal{X}_2,\mathcal{X}_3}$ and $C_{\mathcal{X}_2,\mathcal{X}_5}$, but not $C_{\mathcal{X}_3,\mathcal{X}_5}$). The extra correlation functions are used in iterative imposition algorithms, which will be described in later sections.

At the first glance, the imposition itself is quite ideal, since it can impose full correlations to a TFR at once. It is also a non-iterative procedure, which imposes all the autocorrelation functions and cross-correlation functions obtained from one TFR magnitude to another TFR. However, based on observations, we found that not all the entries in the correlation functions are meaningful, these entries have no perceptual impacts on how human recognize the sound texture. Therefore, it should be reasonable to impose only a selected part of the correlation functions. Furthermore, a perfect imposition may consume too much degree of freedom thus making the generated samples all looks similar. This will also be described in later sections.

4.2 Imposition of Statistical Moments

From the previous section, we know that mean and variance are automatically preserved if autocorrelation function is preserved. Therefore, the remaining moments those to be imposed are skewness and kurtosis. In this section, two ways of imposing the moments are introduced. Both of them are iterative. The first is a traditional gradient projection which is similar to (Portilla and Simoncelli, 2000). The same gradient projection was

employed in (Liao et al., 2013). The first method, which is the temporal domain imposition, do not preserve correlation functions, thus an alternative update is required: impose correlation functions then impose the moments and repeat until the iteration converges. The second method, which is the spectral domain imposition, is more interesting, because it imposes skewness and kurtosis while preserving autocorrelation and cross-correlation functions. While the second method looks superior, the problem is that, the optimization process can only find a local minimum, thus the moment values will not completely equal to the target values, merely approximated values.

4.2.1 Temporal Domain Imposition

The temporal domain here means the temporal axis in the time-frequency domain. We would like to impose moments by changing the sub-band signals directly. A gradient projection algorithm is used to impose the third and fourth moments, the detail of which can be found in (Portilla and Simoncelli, 2000). In the case of skewness η and kurtosis κ, their derivatives with respect to a sub-band signal $\mathcal{X}$ are:

$$\frac{\partial \eta(\mathcal{X})}{\partial \mathcal{X}} \equiv \mathcal{X} \circ \mathcal{X} - \mu_2^{1/2}(\mathcal{X})\eta(\mathcal{X}) - \mu_2(\mathcal{X}) \tag{4.8}$$

$$\frac{\partial \kappa(\mathcal{X})}{\partial \mathcal{X}} \equiv \mathcal{X} \circ \mathcal{X} \circ \mathcal{X} - \mu_2(\mathcal{X})\kappa(\mathcal{X})\mathcal{X} - \mu_3(\mathcal{X}) \tag{4.9}$$

The $\circ$ symbol means the Hadamard product(entry-wise product). Since either finding a single projection function or a set of orthogonal projection functions is difficult, we choose to impose the moments sequentially. We try to approach the target moment with smallest change of signal $\mathcal{X}$ by moving along the gradient's direction. It can be done with the following formula $\mathcal{X}$:

$$\mathcal{X}' = \mathcal{X} + \gamma \frac{\partial f(\mathcal{X})}{\partial \mathcal{X}} \tag{4.10}$$

The target signal $\mathcal{X}'$ can be obtained by solving γ with respect to $f(\mathcal{X}') = m_t$. If there is no real root for γ, a real-value approximation which is closest to the target value will be chosen and continue another projection step. In (4.10), f is either third or fourth moment, where $\frac{\partial f(\mathcal{X})}{\partial \mathcal{X}}$ is the derivative of the moment function respected to $\mathcal{X}$, and m_t is the target value of the desired moment.

With the gradient projection algorithm, we impose the high-order moments to the sub-band signals directly. Unfortunately, the gradient projection would change the correlation functions of the sub-band signals; therefore, an alternative update procedure is required to generate the TFR for the new sound texture sample.

4.2.2 Spectral Domain Imposition

Recall equation (4.7), which is in section 4.1, we talked about another way to understand the procedure. It was said that, the imposition is in fact generating new samples from the original one, by means of adding a phase offset vector to the spectra of original sub-band signals. It means that, if we choose the phase-offset vector properly, we may be able to impose the moments while still preserving all the correlation functions. With the carefully chosen phase offset vector, we could possibly generate a new sound texture sample, which fits all the properties in the statistical descriptions. Therefore, we are going to find an objective function that optimizes all the skewness and kurtosis on all frequency bins respect to the phase offset vector. This moment imposition method is used in the final version of the proposed sound texture synthesis algorithm.

The first step would be figure out how to adjust the skewness and kurtosis in the spectral domain of a signal. This can be found in Appendix A. The appendix also discusses about how to estimate higher order moments in the spectral domain. Now the moment imposition problem can be formulated like this:

Imposing Moments in Spectral Domain:

Given a M-by-N matrix $\mathbf{X}$, which is a TFR of sound texture. The target skewness of each frequency bin is a vector $\eta_t \equiv (\mathcal{M}_\mathcal{X})_3$, the kurtosis of each frequency bin is a vector $\kappa_t \equiv (\mathcal{M}_\mathcal{X})_4$. We would like to find a non-zero row vector θ, which minimizes the objective function f:

$$\arg\min_{\theta} f_1(\theta, \mathbf{X}, \eta_t, \kappa_t) = |\eta(X_\theta) - \eta_t|^2 + |\kappa(X_\theta) - \kappa_t|^2 \tag{4.11}$$

$$\nabla\kappa(X_\theta) \propto \Im(\overline{\mathbf{C}}_{\mathbf{A}_{\hat{\mathbf{X}}_\theta}, \hat{\mathbf{X}}_\theta} \circ \hat{\mathbf{X}}_\theta) \tag{4.12}$$

$$\nabla\eta(X_\theta) \propto \Im(\overline{\mathbf{A}}_{\hat{\mathbf{X}}_\theta} \circ \hat{\mathbf{X}}_\theta) \tag{4.13}$$

$$\nabla f_1(\theta, \mathbf{X}, \eta_t, \kappa_t) \propto (\eta(X_\theta) - \eta_t)\nabla\eta(X_\theta) + (\kappa(X_\theta) - \kappa_t)\nabla\kappa(X_\theta) \tag{4.14}$$

The optimization process finds θ to minimize the least-square error respect to the skewness and kurtosis of each row in $\mathbf{X}$. The gradients are the results derived from appendix

A. According to the appendix, the rows in $\mathbf{X}$ must be standardized before entering the process. The symbols in (4.11) are defined as the following:

$$\mathbf{X} = \begin{Bmatrix} \mathcal{X}_0 \\ \mathcal{X}_1 \\ \vdots \\ \mathcal{X}_{M-1} \end{Bmatrix}, \quad \hat{\mathbf{X}} = \begin{Bmatrix} \hat{\mathcal{X}}_0 \\ \hat{\mathcal{X}}_1 \\ \vdots \\ \hat{\mathcal{X}}_{M-1} \end{Bmatrix}, \quad \mathbf{\Theta} = \begin{Bmatrix} \theta \\ \theta \\ \vdots \\ \theta \end{Bmatrix}, \quad \hat{\mathbf{X}}_\theta = \hat{\mathbf{X}} \circ e^{j\mathbf{\Theta}}$$

Considering rows in $\mathbf{X}$ are real-valued, the rows in $\hat{\mathbf{X}}$ must be Hermitian, thus $\hat{\mathbf{X}}_\theta$ must be also Hermitian about $N_y = \lfloor N/2 \rfloor$:

$$\hat{\mathbf{X}}_\theta = \begin{Bmatrix} \hat{\mathcal{X}}_{0,0}e^{j\theta_0}, & \hat{\mathcal{X}}_{0,1}e^{j\theta_1}, & \cdots & \hat{\mathcal{X}}_{0,N_y}e^{j\theta_{N_y}} & \cdots & \hat{\mathcal{X}}_{0,N-2}e^{-j\theta_2}, & \hat{\mathcal{X}}_{0,N-1}e^{-j\theta_1} \\ \hat{\mathcal{X}}_{1,0}e^{j\theta_0}, & \hat{\mathcal{X}}_{1,1}e^{j\theta_1}, & \cdots & \hat{\mathcal{X}}_{1,N_y}e^{j\theta_{N_y}} & \cdots & \hat{\mathcal{X}}_{1,N-2}e^{-j\theta_2}, & \hat{\mathcal{X}}_{1,N-1}e^{-j\theta_1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \hat{\mathcal{X}}_{M-1,0}e^{j\theta_0}, & \hat{\mathcal{X}}_{M-1,1}e^{j\theta_1}, & \cdots & \hat{\mathcal{X}}_{M-1,N_y}e^{j\theta_{N_y}} & \cdots & \hat{\mathcal{X}}_{M-1,N-2}e^{-j\theta_2}, & \hat{\mathcal{X}}_{M-1,N-1}e^{-j\theta_1} \end{Bmatrix}$$

The objective function (4.11) is a preliminary result. We discovered that there are some problems within by means of experimental investigation. The first problem is that, the weighting between skewness and kurtosis is imbalanced. The kurtosis, due to its fourth order nature, always dominates over the skewness. Even though the skewness is a factor of the lower bound of the kurtosis, thus dependent to the kurtosis to a certain degree, however, the kurtosis still received too much weighting. This problem can be fixed by adding an extra weighting for the skewness, which equals to the length of a sub-band signal, to compensate the imbalance brought by the order difference between the two moments.

The second problem is, the rows in matrix $\mathbf{X}$ are standardized. Therefore, each row receives the same weighting during the optimization. Considering that the optimization process always fall into a local minimum, it is very likely that not all the moments can be fitted. In this case, if we can introduce extra weighting parameters, we can direct the optimization process to raise the importance of some sub-band signals. Another concern is that, we would like to avoid the optimization fall into the so called *ravine condition*(Møller, 1993). The scaled conjugate gradient(SCG) algorithm(Møller, 1993) can mitigate the situation. It uses a quadratic approximation to estimate the Hessian matrix, and use the information to decide the scale of the next step. In order to use the SCG algorithm, the gradient must be exact. The scalar multiplier in the gradient terms thus becomes critical. The gradient with the scalar multiplier is derived in the appendix A. At this time, we can now derive a more powerful version of (4.11), with two extra parameters as the weighting vectors $\mathcal{W}_\eta, \mathcal{W}_\kappa$:

$$\arg\min_{\theta} f_2(\theta, \mathbf{X}, \eta_t, \kappa_t, \mathcal{W}_\eta, \mathcal{W}_\kappa) = \mathcal{W}_\kappa \circ \Delta_\kappa^2 + \mathcal{W}_\eta \circ \Delta_\eta^2 \tag{4.15}$$

$$\Delta_\kappa = \frac{1}{N^4} diag(\overline{\mathbf{A}}_{\hat{\mathbf{X}}} \mathbf{A}_{\hat{\mathbf{X}}}^T) - \kappa_t \tag{4.16}$$

$$\Delta_\eta = \frac{1}{N^3} diag(\overline{\mathbf{A}}_{\hat{\mathbf{X}}} \hat{\mathbf{X}}^T) - \eta_t \tag{4.17}$$

$$\begin{aligned} \nabla f_2(\theta, \mathbf{X}, \eta_t, \kappa_t, \mathcal{W}_\eta, \mathcal{W}_\kappa) = 16 \sum_{i=0}^{M-1} (\mathcal{W}_\kappa \Delta_\kappa)_i \Im(\overline{\mathbf{C}}_{\mathbf{A}_{\hat{\mathbf{X}}}, \hat{\mathbf{X}}} \circ \hat{\mathbf{X}})_{i,j} \\ + \frac{12}{N} \sum_{i=0}^{M-1} (\mathcal{W}_\eta \Delta_\eta)_i \Im(\overline{\mathbf{A}}_{\hat{\mathbf{X}}} \circ \hat{\mathbf{X}})_{i,j} \end{aligned} \tag{4.18}$$

The elements of weighting vectors are determined by the power density of the sub-band signal. Empirically, we found that the square root of the power density is suitable for most of the situations. Therefore, we set the weighting vectors like this:

$$\mathcal{W}_\eta(k) = N_x \sqrt{\sigma^2(\mathcal{X}_k) + \mu^2(\mathcal{X}_k)}, \quad \mathcal{W}_\kappa(k) = \sqrt{\sigma^2(\mathcal{X}_k) + \mu^2(\mathcal{X}_k)} \tag{4.19}$$

Finally, we have come up with a weighted moment imposition method which works in the spectral domain of sub-band signals (4.15). There is at least one global minimum which is the original sound when θ is a zero vector. The method imposes skewness and kurtosis to all sub-band signals by changing the phase offset vector θ. Combined with the non-iterative imposition (4.7) which introduced in section 4.1, it seems to be a compelling method that effectively imposes three kinds of statistical properties while two of them are guaranteed to be the assigned values.

Aside from applying the moment imposition to all sub-band signals together, it can also be applied with a selected set of sub-band signals. In this way, only the cross-correlation functions of the selected sub-band signals are preserved. This can allow the generated signals to produce more variations. Lesser sub-band signals improves the result of imposition, at the cost of some cross-correlation functions destroyed. The trade-off between these two can be adjusted according to the situation. For example, if cross-correlation functions of neighbouring sub-band signals are more perceptually important than the others, a good idea will be including the neighbouring sub-band signals to preserve these cross-correlation functions.

Unfortunately, combining (4.7) and (4.15) creates a new problem. For some sound textures, about slightly less than half among the samples, the algorithm has a high chance to generate a TFR that is merely the original TFR with a random delay. The

generated TFR is different from the original in detail, but the global structure is the same as a delayed version of the original TFR. There are several possible causes. Since we have tried the initialization of θ in many different ways, this is not likely due to bad initial values. One possibility is that the solution plane has a global tendency which acts like a steep valley, and most of the time the optimization process steps right down into the bottom of the valley, which is the global minimum. Another possibility is that, there are simply too few minimum points other than the global minimum. Anyway, the phenomenon suggests that we might impose too many properties such that there's no degree of freedom left for other variations. Moreover, it is possible that, the full set of correlation functions along with moments in each sub-band has the potential to determine the whole signal. A work related to this situation is (Yellott Jr et al., 1993). Yellott's triple correlation uniqueness(TCU) theorem indicates that, using all the third-order statistics can uniquely determine a finite-size monochrome image. However, we do not use all the third-order statistics, thus further investigation would be required to clarify the situation.

In order to increase the degrees of freedom and allow more variations when generating new sound texture samples. We would like to seek a way to apply only a selected part of a cross-correlation function. This will be described in the next section.

4.3 Partial Imposition of Correlation Functions

In this section, we would like to develop a method to apply only a selected part of the correlation functions. Some parts of the correlation function may not be perceptually important. For example, those parts which are far from the center. It is less likely to have a meaningful correlation across such a long time span. The final version of the proposed sound texture synthesis uses this method as the correlation imposition algorithm.

If we would like to impose only a part of the correlation, the method we proposed in section 4.1 is no longer applicable. The method only assures the equivalence of correlation functions between the original and the generated sound textures under the conditions that:

1) The sub-band signals of the generated sound texture have the same magnitude spectra as the sub-band signals of the original sound texture sample.

2) When the first condition holds, the phase differences of sub-band signal spectrum are the same between the original and generated sound textures.

These conditions cannot hold when we want to apply only a part of the correlation functions. Therefore, without further investigation or new discoveries, it would be necessary to fall back to the traditional iterative optimization. However, we would like to keep using the phase differences to modify the cross-correlation functions, thus we will assume that the magnitude spectrum is fixed but not necessarily the same magnitude spectrum as the original sound. We can begin with the common least-square error optimization. Consider an arbitrary M-by-N TFR, $\mathbf{X}$, we would like to impose the cross-correlation functions of the original sound texture, $\mathcal{C}$, on $\mathbf{X}$. As mentioned in section 3.3, only a selected set of cross-correlation functions will be taken into account. We use a symmetric matrix $\mathbf{V}$ to indicate this. If a cross-correlation function $\mathcal{C}_{\mathcal{X}_i,\mathcal{X}_j}$ is selected, then $(\mathbf{V})_{i,j} = v_{i,j} = 1$.

$$\mathbf{V} = \{v_0, v_1, \cdots, v_M\}^T \tag{4.20}$$

$$v_{i,j} = \begin{cases} 1, & |i-j| \bmod 3 \equiv 0 \,\text{or}\, |i-j| = 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.21}$$

The objective function, ϕ_i, respect to the ith row of $\mathbf{X}$ is:

$$\arg\min_{\theta_i} \phi_i(\theta_i, X_i, v_i, \mathcal{C}) = \sum_{j \in \{j : v_{i,j}=1\}} \Big(\sum_{\tau=-\infty}^{\infty} |C_{X_i^\theta, X_j}(\tau) - \mathcal{C}_{\mathcal{X}_i,\mathcal{X}_j}(\tau)|^2 \Big) \tag{4.22}$$
$$X_i^\theta = \mathcal{F}^{-1}\{|\hat{\mathcal{X}}| e^{j\theta_i}\}$$

(4.22) evaluates the sum of all least square error of cross-correlation functions related with X_i^θ. X_i^θ is the ith sub-band signal of the generated sound texture which the phase is controlled by θ_i. Begin from (4.22), now we can discuss about how to impose only a part of the cross-correlation function to $\mathbf{X}$.

A possible pitfall about imposing partial cross-correlation function is that: *If some parts are irrelevant, simply consider them as zero.* The action of enforcing irrelevant terms to zero is just imposing another value; the terms will still contribute to the objective function. Setting the terms with randomly small values does not work either. Assigning another value is just forcing the generation process to generate another specific sample, not generating a sample from the set of possible samples of the sound texture. The statistical description should only contain properties that are relevant to the perceptual recognition of a sound texture, not the properties of a specific sound texture sample. Besides, cross-correlation functions with arbitrarily assigned values can be inconsistent,

as described in 4.1. A feasible solution should achieve: *Regardless how these irrelevant terms change, the value of objective function is unaffected.*

The solution we propose here is to include a window-shaped weighting vector. The weighting vector diminishes for the further part of the correlation function and eventually reaches zero. Depending on the shape of the vector, the weighting vector can reduce or remove the influence of irrelevant terms, making them contribute nothing to the objective function. The vector is denoted as w. The value of elements in w is always between 0 and 1. For example, Tukey window(Harris, 1978) window can be a good choice of w, as it is flat around the center then gradually decreases till the edge. The objective function with the weighting vector w is:

$$\begin{aligned} \arg\min_{\theta_i} \phi_i^w(\theta_i, X_i, v_i, \mathcal{C}) &= \sum_{j\in\{j:\, v_{i,j}=1\}} \Big(\sum_{\tau, w(\tau)\neq 0} w^2(\tau)(|C_{X_i^\theta, X_j}(\tau) - \mathcal{C}_{\mathcal{X}_i,\mathcal{X}_j}(\tau)|)^2 \Big) \\ X_i^\theta &= \mathcal{F}^{-1}\{|\hat{\mathcal{X}}|e^{j\theta_i}\} \end{aligned} \tag{4.23}$$

In order to reduce redundant computation, we can make some observations on (4.23). We found that these matrices are constant during the optimization process:

$$\begin{aligned} \mathbf{W}_i^2 &= \{\overbrace{w^2, w^2, \cdots, w^2}^{\sum_j v_{i,j}}\}^T \\ \mathbf{B}_i &= \{|\hat{X}_i| \circ \overline{\hat{X}}_j, : \forall j,\, v_{i,j} = 1\}^T \\ \mathbf{C}_i &= \{\mathcal{C}_{\mathcal{X}_i,\mathcal{X}_j}, : \forall j,\, v_{i,j} = 1\}^T \end{aligned} \tag{4.24}$$

$\mathbf{W}_i^2$ is a row-major matrix with rows tiled with the squared weighting vector, w^2. The number of rows in $\mathbf{W}_i^2$ is equal to $\sum_j v_{i,j}$, which is the number of selected cross-correlation functions. $\mathbf{W}_i^2$, $\mathbf{B}_i$ and $\mathbf{C}_i$ have the same size. If all the matrices in (4.24) are pre-calculated, (4.23) and its derivatives can be evaluated with ease:

$$\begin{aligned}\Theta &= \{\overbrace{\theta_i, \theta_i, \cdots, \theta_i}^{\sum_j v_{i,j}}\}^T \\ \tilde{\mathbf{C}}_i &= C_{X_i^\theta, X_j} = \mathcal{F}^{-1}\{\mathbf{B}_i \circ e^{j\Theta}\} \\ \Delta\phi_i &= \tilde{\mathbf{C}}_i - \mathbf{C}_i\end{aligned} \tag{4.25}$$

$$\begin{aligned}\phi_i^w(\theta_i, X_i, v_i, \mathcal{C}) &= \sum_{\forall(p,q)} (\mathbf{W}_i^2)_{p,q} \circ (|\Delta\phi_i|^2)_{p,q} \\ \nabla\phi_i^w(\theta_i, X_i, v_i, \mathcal{C}) &= 4\sum_q \left(\Im\{\overline{\mathcal{F}}\{\Delta\phi_i \circ \mathbf{W}_i^2\} \circ \tilde{\mathbf{C}}_i\}\right)_{p,q}\end{aligned} \tag{4.26}$$

(4.24) to (4.26) describes how to adapt the cross-correlation between X_i and its selected neighbours in v_i by changing the phase of $\hat{X}_i$. We again use the scaled conjugate gradient(Møller, 1993) algorithm to minimize ϕ_i. This process will go through every row in $\mathbf{X}$ in a round-robin manner as (4.27) for several times. This method can also be with the spectral moment imposition in section 4.2.2. The spectral moment imposition changes only the phase-offset vector, thus the phase differences between sub-band signals are preserved. Therefore, after this partial cross-correlation function imposition, the spectral moment imposition can be applied without destroying the correlation functions. These two methods are both used in the final version of the sound texture synthesis procedure.

$$\arg\min_{\theta_i} \phi_i^w(\theta_i, X_i, v_i, \mathcal{C}), \quad \forall\, 1 \le i \le \left\lfloor \frac{M-1}{2} \right\rfloor \tag{4.27}$$

It is tempting to extend (4.23) such that it can change all the phases in $\mathbf{X}$ together to adapt all the selected cross-correlation functions in $\mathbf{V}$. However, the experiment result with this extension is less satisfying. The reason is that the extended version creates an artifact, which is a huge vertical line in the spectrogram, sounds like a click. A possible reason is that when all the phases are changed together, in the beginning of the process, we found that the gradient has the tendency to guide all the phases to be more aligned with other bands. This leads the optimization process to a non-feasible local minimum. This phenomenon also appears if someone uses a 'shallow' optimization strategy when using (4.23). The shallow optimization means that it gives each X_i a very low maximum iteration number such that the SCG always stops prematurely, but raises the number of round-robin cycles. For example, if a standard optimization is 10 round-robins with 100

iterations for each row, the shallow optimization is 100 round-robins with 10 iterations for each row. The total steps are equal between these two, which is 1000, but the shallow optimization gives less satisfying results.

Another possible extension of this method is to allow the magnitude spectrum to be changed together with the phases, thus enabling the imposition of partial autocorrelation functions. The difficulty of this extension is that, if both the magnitude spectrum and phases are allowed to change freely, it becomes an optimization respect to a complex valued vector. The complex derivative of the objective function may not exist, in other words, not complex differentiable. The problem is discussed in appendix B. This is a limit for those who impose statistics in the spectral domain of sub-band signals.

4.4 Discussion

Moments	Iterative	Preserve Correlation	Accuracy
sect. 4.2.1	Yes	No	Usually exact value
sect. 4.2.2	Yes	Yes	Approximated value
Correlation Functions	Iterative	Preserve Moments	Accuracy
sect. 4.1	No	No	Exact value
sect. 4.3	Yes	No	Approximated value

TABLE 4.1: *The comparison between the imposition methods introduced in this chapter.*

This chapter we introduced two methods to impose the moments and two methods to impose the correlation functions. Imposing moments directly on the sub-band signal in the temporal domain (section 4.2.1) usually can adapt the moments to exactly fit the target value, but has a side effect that will produce unwanted changes to correlation functions. In contrast, the spectral domain imposition(section 4.2.2) of moments may not perfectly adapt the moments to the target value, but will preserve the correlation functions. It is interesting that, imposing full correlation functions(section 4.1) requires no iteration and the accuracy of the imposition is best. However, it imposes too much information and takes away too much degree of freedom. In the other hand, even though the partial imposition of correlation functions(4.3) is an iterative process and usually can only achieve an approximated imposition, it is still more favourable than the full imposition method. The comparison is shown in table 4.1.In the preliminary stage of the research, we have tried the combination of temporal domain moment imposition(sect. 4.2.1) and full correlation function imposition (sect 4.1) in (Liao et al., 2013). Nevertheless, after several experiments, we found the best combination in general is the spectral moment imposition(sect. 4.2.2) combined with partial correlation function imposition(sect. 4.3), which yields high quality result but also maintained the variety of generated samples.

Chapter 5

Proposed Method, Summary

The proposed method is an analysis-synthesis scheme, which creates a statistical description based on an input sound texture then generates resynthesized sound texture samples from the statistical description. The statistical description contains the magnitude statistics of every frequency bins in the time-frequency domain. The statistics of frequency bins include autocorrelation functions and moments of the magnitude. Besides the statistics of individual frequency bins, the statistical description also includes cross-correlation functions between some frequency bins. The synthesis begins from a TFR of a randomly generated noise. The amplitude spectra of the sub-band signals of this generated noise are replaced by the spectra of the autocorrelation functions. Subsequently, the correlation functions and moments are imposed. In the last step, the phase of the TFR is reconstructed and the new sound texture is resynthesized. In this chapter, we would like to give a complete description of the whole analysis synthesis scheme. Implementation details and parameter settings are also discussed.

5.1 Analysis

The analysis part takes a sound texture as input, follow the steps described in section 3.3 to obtain the statistical description $\Phi = \{\mathcal{M}, \mathcal{A}, \mathcal{C}\}$. Currently the underlying TFR is the STFT. The analysis window of the STFT is a 128ms periodic Hanning window. The hop size between neighbouring analysis frames is 8ms. Larger window size gives better representation of low frequency perceptual bands, but needs more computation on cross-correlation functions. Empirically, the suitable size of Fourier transform should between from $64ms$ to $256ms$ and not shorter than the window size. In our case, the preferred choice is $128ms$.

5.2 Synthesis

The synthesis takes a statistical description $\Phi = \{\mathcal{M}, \mathcal{A}, \mathcal{C}\}$ as input. From this statistical description, we can generate arbitrary long sound texture samples. These new samples will all have the same statistic as described in Φ. The synthesis procedure contains four stages before the final resynthesize; these steps are described in the following subsections. If a new sample that is longer than the original is required, the new samples will be generated block-by-block; the size of a single block is the same as the original.

5.2.1 Initialization, Preprocessing

The initialization stage generates a randomized TFR $\tilde{\mathbf{X}}$. In our case it is the STFT spectrogram of a random noise. Due to the symmetrical property of the STFT spectrogram, only half of the frequency bins are used (including the DC term and the Nyquist frequency). The magnitude of the STFT spectrogram will be compressed as described in equation (3.5), $\tilde{\mathcal{X}}_k(n) = |\tilde{X}(n,k)|^{2c}$. After the magnitude is compressed, the moment of each sub-band signal will be directly adjusted according to the statistical description Φ. In fact, different types of the noise will slightly induce different tendencies to the resulting sound texture. For example, if the noise is a randomized impulse, the resynthesized fire texture has a higher tendency to have more clicks. However, in the general case, we use Gaussian noise to be the random noise. If a longer sample is required, the initialization generates all the blocks at once, and then procedure repeats on different blocks sequentially.

5.2.2 Correlation Function Imposition

In this stage, the sub-band signals of $\tilde{\mathbf{X}}$ will be Fourier transformed and their spectral magnitude will be replaced with those in Φ. This imposes autocorrelation functions to $\tilde{\mathbf{X}}$, just like (4.4).

$$\hat{\tilde{\mathcal{X}}}_k \leftarrow |\hat{\mathcal{A}}_k| e^{j\theta(\hat{\tilde{\mathcal{X}}}_k)} \tag{5.1}$$

Then, the partial cross-correlation imposition(section 4.3, 4.27) is applied to each sub-band signal once in a round-robin manner. The cross-correlation selection is the same as section 4.3. The weighting vector w is a 2048-point rectangular window. Considering the hop size, the window spans around $\pm$0.819ms time scale. The effect of different length of weighting vector will be shown in chapter 6. The error improvement of cross-correlation

functions is recorded in the dB scale and used as a metric to determine whether the whole process has converged. If the average improvement of cross-correlation error is smaller than a threshold of 3dB, then the process stops, proceeds to the next block. Otherwise, the process executes 4.27 again. This can repeat up to 10 times, which is the maximum limit. If all blocks have been processed, then the process enters the next stage.

5.2.3 Moment Imposition

The moment imposition method in section 4.2.2 is applied to the whole $\tilde{\mathbf{X}}$. After the moment imposition is done, the process proceeds to the next block. If all blocks blocks are finished, then the process enters the next stage.

It was mentioned in the beginning of chapter 4 that the proposed algorithm does not preserve the non-negativity of sub-band signals. It is hard to preserve non-negativity when we alter the sub-band signal in its spectral domain. The amplitude compression helps to mitigate the problem, since the amplitude compression will enlarge the values that were originally between 0 and 1. However, sometimes there are still few negative values appear. The negative values of sub-band signals are replaced by zeros before the next stage.

5.2.4 Phase Reconstruction

After all the blocks have been generated, the amplitude compression of the TFR is restored. Then, we use LeRoux's algorithm(Le Roux et al., 2008, Le Roux et al., 2010) to reconstruct the phases of the spectrogram. To be mentioned, the phases are reconstructed in such a way that neighbouring analysis frames have no inconsistencies in their overlap part in time domain. The reconstruction algorithm has no assumption about how the phase changes in the original signal. It uses a trimmed time-frequency convolution to reconstruct the phases. According to LeRoux's suggestion, it states that the reconstruction should begin from coefficients with large magnitude in order to place inconsistencies in the smaller coefficients. However, sorting all the coefficients according to their magnitudes is time consuming, especially if the generated spectrogram is long. In contrast, we used a threshold value that is initially equals to the largest magnitude. The reconstruction is then sequentially applies to coefficients that are larger than the threshold. After all the coefficients are processed, the threshold is reduced by half and the process repeats. This is a simplified alternative than sorting the coefficients. Coefficients that are 72dB smaller than the initial threshold are omitted, their phases are randomly assigned. Empirically, it takes 40 to 60 iterations to get a high quality result.

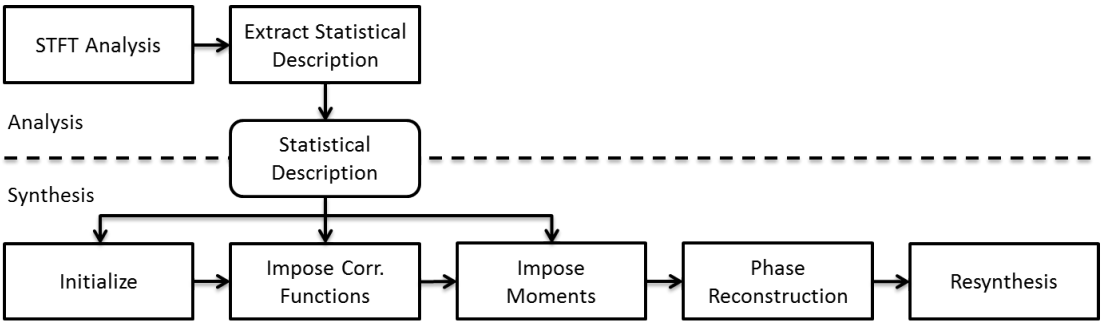


FIGURE 5.1: *The workflow overview of the proposed analysis-synthesis scheme*

However, the phase reconstruction cannot remove all the inconsistency. It is possible that the generated TFR magnitude by the proposed algorithm has inherent inconsistency, which cannot be resolved by the phase reconstruction. It is also possible that the phase reconstruction unluckily falls into a bad local minimum. Therefore, the resynthesized signal will not have exactly the same spectrogram as the generated TFR.

After the phase reconstruction is done, an inverse STFT is applied and the signal waveform of the new sound texture sample is obtained.

5.3 Discussion

In this chapter, we have disclosed the full analysis and resynthesis workflow. The analysis synthesis scheme we proposed here only relies on the statistical description to generate new samples of sound textures. The auto-correlation function is preserved by the method in section 4.1. The stage of correlation function imposition imposes partial cross-correlation until the improvement is smaller than a threshold. Besides, it is possible to use a weighting vector with different shapes for the partial cross-correlation imposition. The workflow is depicted in fig 5.1.

Chapter 6

Evaluation

In this chapter, we evaluate the proposed algorithm in different ways. They can be categorized as objective and subjective. The objective evaluation provides numerical information about the resynthesized result and the profile of the algorithm itself. This includes the algorithm's run time, comparing the correlation functions and moments in terms of SNR(signal-to-noise ratio) and some statistics about the efficiency of the optimization process. The subjective evaluation examines the resynthesized result by human perception. One of the experiments instigates the perceptual difference of different cross-correlation function length. The other two experiments compare the proposed algorithm with McDermott's(McDermott and Simoncelli, 2011a) and Bruna's work(Bruna and Mallat, 2013). The sound samples used in these evaluations was obtained from McDermott's website(McDermott and Simoncelli, 2011b). There are a total of 15 samples; each of them lasts from 5 to 7 seconds. We also gladly received the matlab implementation of McDermott's algorithm from him. Thanks to his kindness.

6.1 Objective Evaluation

This section contains two parts, the profiling of algorithm and the quality measurement of the synthesized result. The profiling of the algorithm shows the bottleneck of the algorithm and provides information for future optimization when implementing with other programming languages. The quality measurement investigates the SNR of the statistics of the synthesized sound samples. However, the SNR of the statistics can only serve as a reference, since the perceptual quality has no significant observable link with the SNR.

Stage Name	Time Consume%	Gradient	Objective Function
Partial CCF Imposition	69.4%	44.6%	35.1%
Moment Imposition	24.4%	65.5%	3.8%
Phase Reconstruction	5.2%	n/a	n/a
Others	1.0%	n/a	n/a

TABLE 6.1: *The profiling result of the proposed algorithm. The second column is the time consumption of the stage. The third column is how much time that stage spent in calculating its gradient. The fourth column is how much time that stage spent in calculating the objective function value.*

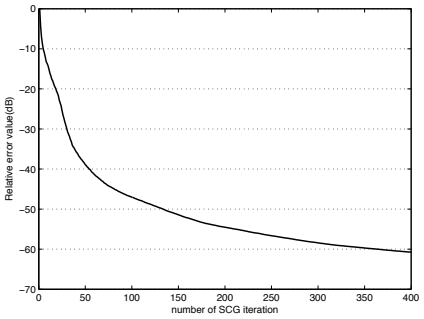


FIGURE 6.1: *The relative error after each SCG iteration of spectral domain moment imposition.*

6.1.1 Profiling

In this section, we investigate the efficiency of the proposed algorithm. In order to obtain an estimate, the algorithm is profiled by the MATLAB profiler. The averaged profiling result can be seen in table 6.1.

According to table6.1, the most time-consuming process is the partial cross-correlation function imposition(69.4%). The process spends 79.7% self-time calculating the gradient and the value of objective function. The second most time-consuming process is moment imposition(24.4%). It takes 69.3% self-time in calculating gradient and objective function. The phase reconstruction takes only 5.2% of time.

The average running time to generate a high quality sample is 26 minutes. It is slower than the 10-minute average of an earlier version of the algorithm(Liao et al., 2013) due to the introduction of partial cross-correlation function imposition. It is still an improvement as McDermott's algorithm(McDermott and Simoncelli, 2011a) take 69 minutes on average to generate a new sample of the same texture on the same machine.

Fig. 6.1 depicts the decreasing of the relative error after each SCG iteration. It usually takes around 130 iterations to reach $-50dB$ and aound 360 iterations to reach $-60dB$. In the practical case, the iteration stops after the error value has been reduced by $60dB$.

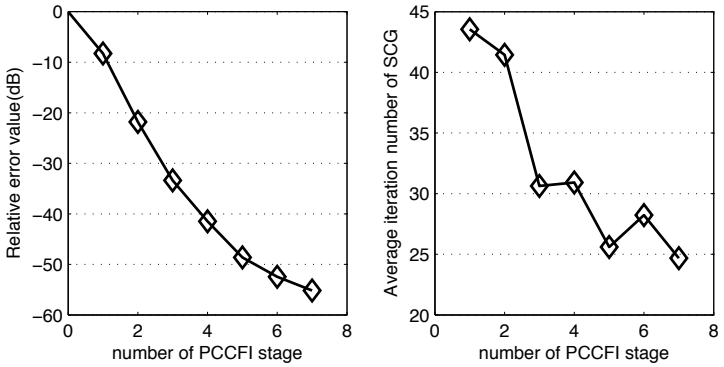


FIGURE 6.2: *Left: The relative error after each partial cross-correlation function imposition(PCCFI) stage. Right: Average steps required to reach the local minimum of* Φ_i^w *in each PCCFI stage. One PCCFI stage means one full round-robin of (4.27).*

Further analysis on the partial cross-correlation function imposition reveals how the error value of cross-correlation function decreases after each full round-robin of sub-band signals. Fig. 6.2 shows the result. It shows that the relative error decreases to -40dB of the initial value after the fourth round-robin. The left sub-figure depicts how the total error value of objective function decreases after each round-robin of Φ_i^w. The right sub-figure depicts how many iteration of SCG required to reach the local minimum of Φ_i^w. Although the number of iteration has a trend to decrease, but still vary between 45 and 25. This may suggests that the final threshold can be raised a little to reduce the computation time at the cost of a small loss of quality.

6.1.2 Measurement of Statistics of Resynthesized Sounds

In this section, we evaluate the statistics of the resynthesized signal, and compare the statistics to those in the statistical descriptions. The statistical descriptions were obtained from the original sound textures. The resynthesized signal is analysed with the procedure which extracted the statistical description(section 3.3). After the TFR of the resynthesized signal is obtained, we evaluate the autocorrelation function and moments for each sub-band signal. In the mean while, we also evaluate cross-correlation functions for all sub-band signal pairs. These statistics are then compared with those in the statistical description, evaluating the SNR. The SNRs of cross-correlation functions measure only those parts preserved during the partial cross-correlation function imposition. The result is shown in table 6.2. This is a result averaged over all the sound samples. Mean and variance are omitted since they are a part of the autocorrelation function. Although the error has been reduced by more than 50 dB as shown in Fig. 6.2, the resynthesized SNR still around 20dB. One reason is that, as described in chapter 4, the non-negativity is not strictly preserved. There are also inconsistencies that cannot be solved by the phase reconstruction, as mentioned in section 5.2.4.

	SNR(dB)
ACF.	18.8252
CCF.	21.0343
Skewness	23.2046
Kurtosis	11.2814

TABLE 6.2: *The average Signal-to-Noise Ratio(SNR) of the resynthsized sound. The SNRs of cross-correlation functions measure only those parts preserved during the PC-CFI. The band-wise SNR of each sound texture can be found in appendix C.*

	Signal-to-Noise Ratio (dB)			
	ACF	CCF	Skewness	Kurtosis
Bubbling water	16.7199	23.8033	26.2168	11.6484
Babbling	12.0696	10.8632	26.8544	19.3550
Pneumatic drills	19.7754	20.1961	21.0614	1.2464
Applauses	21.2986	23.5977	30.6645	12.1657
Wind whistling	21.6987	18.5240	28.5911	18.5077
Bees	24.6775	26.7752	17.9864	15.1675
Helicopter	13.6598	26.8988	25.9622	22.0590
Sparrows	24.5382	27.8615	33.3221	19.8786
Heavy rain falling	27.3821	30.6388	18.1909	1.1900
Steam railroad	16.0533	14.5106	21.7264	15.9459
Fire crackling	14.0927	14.2273	29.1970	14.7006
Insects in a Swamp	23.7160	27.3628	16.8559	1.9868
Rustling papers	14.3702	11.6335	22.8127	13.7644
Lapping waves	9.9048	14.8980	8.0561	2.3033
Stream near waterfall	23.7241	22.0733	20.5716	-0.6979

TABLE 6.3: *The detailed Signal-to-Noise Ratio(SNR) of resynthsized sound samples. The SNRs of cross-correlation functions measure only those parts preserved during the PCCFI.*

The SNR of statistics of each file is shown in table 6.3. The proposed algorithm prioritizes in optimizing correlation functions over the moments; moments are imposed under the condition that correlations are preserved. It can be the reason why the kurtosis is sometimes not well fitted. However, the bad fitting of kurtosis does not pose significant influence in the perceptual score of sound textures. These textures do not receive a specifically bad perceptual scoring.

6.2 Subjective Evaluation

The subjective evaluation is conducted with a form which is adapted from the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor)(ITU-R, 2014) test. Each test case contains a hidden reference and an anchor. The hidden reference is a segment cut from the original sound. The anchor serves as a base-line standard, usually a low-quality sound. The standard suggests an anchor, which is an original sound filtered with

3.5khz-lowpass filter. This is not suitable for the perceptual test of sound textures. The perceptual test of sound texture is not aimed to determine whether two sounds are the same, but to determine whether two sounds belong to the same sound scene. Therefore, the anchor used in the experiment is the proposed algorithm with a relatively short cross-correlation function($\pm 204.8ms$). It has to be clarified that, although the format of the test is similar to the MUSHRA test, it is not a standard MUSHRA test as stated in the (ITU-R, 2014), due to the fundamental difference between the quality measurement of audio codecs and the perceptual test of sound textures. Additionally, according to (Zielinski et al., 2007), the form of MUSHRA test has an inherent bias which will affect how people rate a specific sound sample according to the accompanying samples in the same test case. Therefore, comparing the scores of a sample between different experiments is meaningless.

In these perceptual tests, we asks the participants to rate the sounds according to two different criteria, *Quality* and *Alikeness*. The quality includes whether a sound contains artifact and whether the temporal movement is natural. The alikeness measures whether the sounds are perceptually belong to the same sound scene, or, in other words, belong to the same sound texture.

There are three experiments. The first experiment investigates how the length of cross-correlation function affects the sound texture perception of the auditory system. The second experiment is a comparison of the proposed algorithm to the work of (Bruna and Mallat, 2013). The third experiment compares the proposed algorithm with McDermott's workMcDermott and Simoncelli (2011a).

The three tests are arranged in a similar format. One experiment contains 15 test cases. Each test case contains 5 sound samples. The first one is the original and explicitly marked as the reference. The rest are randomly positioned for each participant. The four samples include one hidden reference and one anchor, the rest are the samples resynthesized by the algorithms to be compared. The exception is Experiment 1, which includes one hidden reference and three synthetic results to be compared. The participant can play the sound in arbitrary order and arbitrary many times. The participants are instructed to rate the samples according to quality and alikeness. Quality measures how much artifact in the test samples, and alikeness measures the perceptual similarity between the samples and the original. There were two tables(6.4, 6.5) presented to the participants with questions regarding to the quality and alikeness. Following these questions, there are descriptions explaining the mapping between the adjectives and the continuous quality scales(CQS:$0-100$).

The perceptual test suggests the participants to use headphones. In the bottom of the test page, the participants are asked if they were using headphones. Some participants

Quality	
How do you feel about the quality of the sound sample?	
0 – 19	Bad, heavy artefact, very annoying.
20 – 39	Poor, moderate artefact, annoying.
40 – 59	Fair, minor artefact, slightly annoying.
60 – 79	Good, small artefacts, but not annoying.
80 – 100	Excellent, minimal or no artefacts.

TABLE 6.4: *The rating standard of Quality.*

Alikeness	
Does the sound sample perceptually belongs to the same physical scene as the original?	
0% – 19%	Unlikely
20% – 39%	Might be
40% – 59%	Maybe
60% – 79%	Probably
80% – 100%	Certainly

TABLE 6.5: *The rating standard of Alikeness.*

are researchers who worked in the related field, they are marked as expert participants. The number of participants in each perceptual experiment is listed in table 6.6. It is worth mentioning that all the expert participants were using headphones during the perceptual test.

	Total	Expert	Non-Expert	Headphone	w/o Headphone
Experiment 1	10	6	4	10	0
Experiment 2a	13	6	7	13	0
Experiment 2b	24	15	6	21	3

TABLE 6.6: *The number of participants in each experiment.*

The results of perceptual tests are analysed with the Student's t-test of non-equal variances(a.k.a Welch's test, 6.3 , (Welch, 1947)) under the significance level $p = 0.05$. Since the comparison is usually made between two groups, a two-sample Student's t-test is sufficient. The confidence intervals in Fig. 6.3, Fig. 6.4 and Fig. 6.5 are evaluated by the one-sample t-test, which is the 95% confidence interval.

The pdf of t-distribution is: (Γ is the gamma function)

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$
$$t_{p/2,n} = f^{-1}(p/2, n) \tag{6.1}$$

The confidence interval of $p = 0.05$:

$$\left(\mu(x) - t_{p/2,n}\frac{\sigma(x)}{\sqrt{n}}, \mu(x) + t_{p/2,n}\frac{\sigma(x)}{\sqrt{n}}\right) \tag{6.2}$$

The two-sample t-test, non-equal variance (Welch's t-test)

$$t = \frac{\mu(x) - \mu(y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \tag{6.3}$$

There are several feedbacks from the participants of the experiment 2a, indicating that the samples were too short thus being hard to determine the quality and alikeness. Nonetheless, it seems that most of the participants can still determine the quality and alikeness with a short sample.

6.2.1 Experiment 1: The effect of different cross-correlation function length

The cross-correlation length in the partial cross-correlation function imposition is an important parameter. We know that auditory system can sense the difference of spectro-temporal correlations across an intermediate timespan(Depireux et al., 2001, Mesgarani et al., 2009). It is interesting to know whether the length of cross-correlation function plays an important role in determining the quality/alikeness of sound textures. Therefore, this experiment compares the effect of different lengths of cross-correlation functions. Three different lengths are used for the cross-correlation functions: $\pm 204.8ms$, $\pm 409.6ms$ and $\pm 819.2ms$. The three resynthesized samples and one hidden reference thus form a test case. The result of the total 15 test cases is shown in Fig. 6.3.

There are some sound textures do not benefit from longer cross-correlation functions, such as *Bubbling water*, *Pneumatic drills* and *Heavy rain falling*. These sound textures generally have plentiful brief wide-band events and lack of structure that persist over a longer time span. It means that these textures can be properly synthesized with a shorter cross-correlation function length. The suitable cross-correlation function length depends on the nature of different sound textures. However, it is still important to find a length that is suitable for most of the sound textures.

In the result, 8 out of 15 files shows that the quality and alikeness score are proportional to the length of cross-correlation function. The same trend also appears in the overall averaged result in Fig. 6.3(c) and Fig. 6.3(d). It seems that the length of cross-correlation function does affect the perceptual feeling to the texture. The Student's t-test(Welch, 1947) shows that there's a significant difference of mean quality and alikeness

score between long CCF and medium CCF. However, the difference of perceptual score is not statistically significant between medium CCF and short CCF. It seems that the length difference between the long CCF($\pm 819.2ms$) and the medium CCF($\pm 409.6ms$) is relevant to the auditory system, thus the perceptual score is significantly higher.

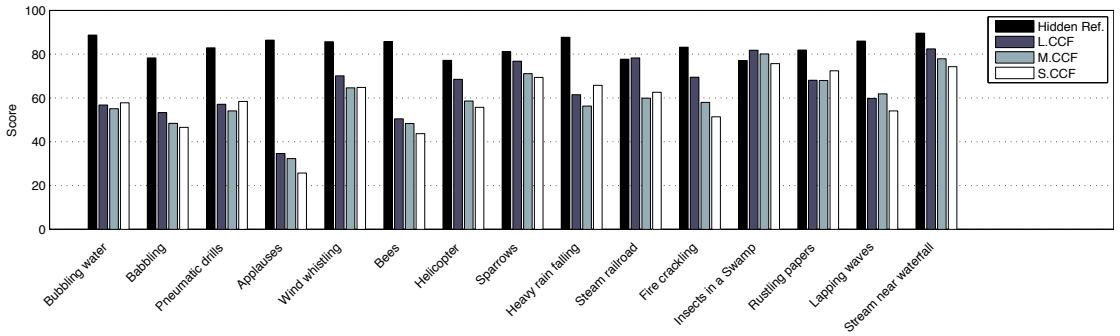
6.2.2 Experiment 2a: Compare with Bruna's work

In this experiment we compare the proposed algorithm with Bruna's work(Bruna and Mallat, 2013), the result is depicted in Fig. 6.4. His work uses scattering moments to characterize sound textures. One goal of his work is to synthesize sound texture with a small set of coefficients. The intent of this comparison is to know how the proposed algorithm performs compared with the scattering transform. As can be seen from Fig. 6.4, due to the vast amount of parameter coefficients in the proposed statistical description, the proposed algorithm outperforms Bruna's work in the overall score. Despite of that, in some textures, Bruna's algorithm performs roughly on par with the proposed algorithm. Considering how small his parameter set is, it means that the scattering moment should be a competitive option as a sparse signal representation for sound textures.

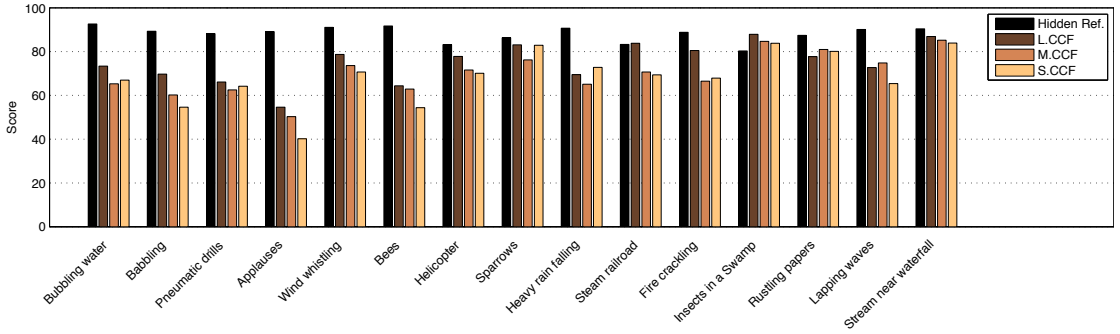
6.2.3 Experiment 2b: Compare with McDermott's work

In this experiment, the proposed algorithm is compared with McDermott's work(McDermott and Simoncelli, 2011a), the result is depicted in Fig. 6.5. McDermott's work uses moments and correlation functions with a perceptual frequency scale filter bank to characterize sound textures. Since the proposed algorithm works with linear frequency scale and reconstructs phases from the magnitude of spectrogram, it would be interesting to know whether the proposed algorithm can achieve the same perceptual quality and alikeness as McDermott's algorithm.

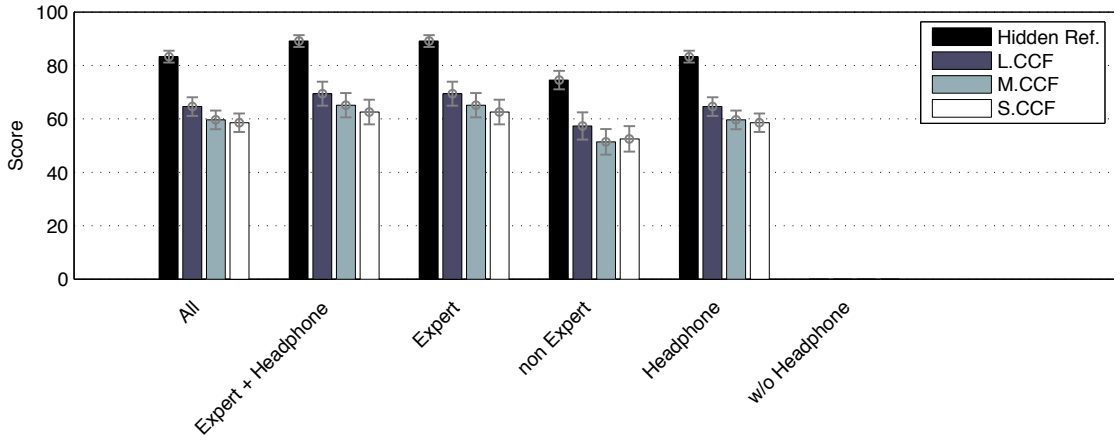
The experiment shows that the proposed algorithm and McDermott's algorithm has no significant difference in terms of overall quality and alikeness. However, there are some sound samples favours in one algorithm and some other samples favours in another algorithm. The difference may came from that, the proposed algorithm prioritizes in imposing correlation functions. In contrast, McDermott's algorithm imposes moments in both the perceptual bands and the modulation bands of each perceptual band. In cases that the most perceptual important statistics are the moments for some sound textures, it is reasonable that McDermott's algorithm performs better in the perceptual test, and, vice versa. Although not being statistically significant ($p = 0.0557$), still, it seems that McDermott's algorithm is more favoured in terms of alikeness. This experiment shows



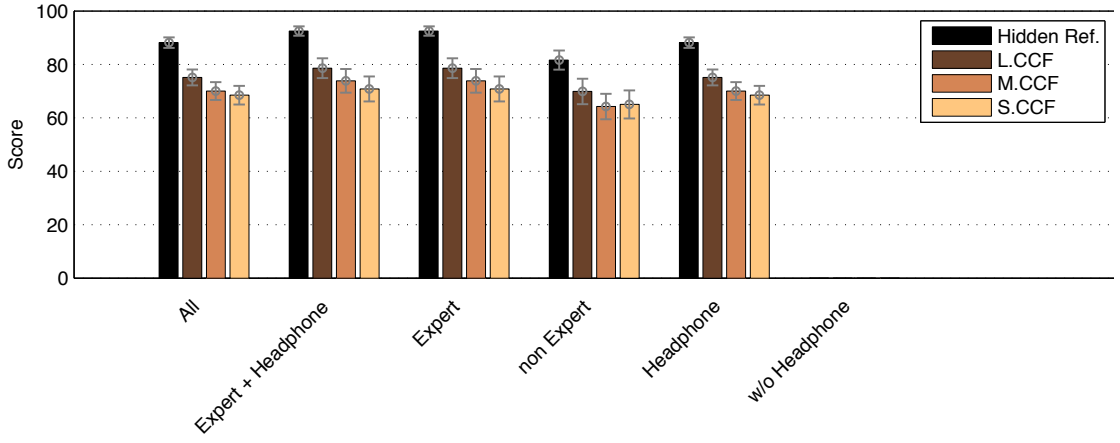
(A) *The quality score of individual files*



(B) *The alikeness score of individual files*



(C) *The quality score averaged over all files(errorbar: 95% confidence interval)*



(D) *The alikeness score averaged over all files(errorbar: 95% confidence interval)*

FIGURE 6.3: *The result of experiment 1 . Hidden Ref: Hidden reference, S.CCF:* $\pm 204.8ms$*, M.CCF:* $\pm 409.6ms$*, L.CCF:* $\pm 819.2ms$

that, the linear frequency scale signal representation is capable of reaching the same quality as the perceptual based signal representation with a shorter computation time.

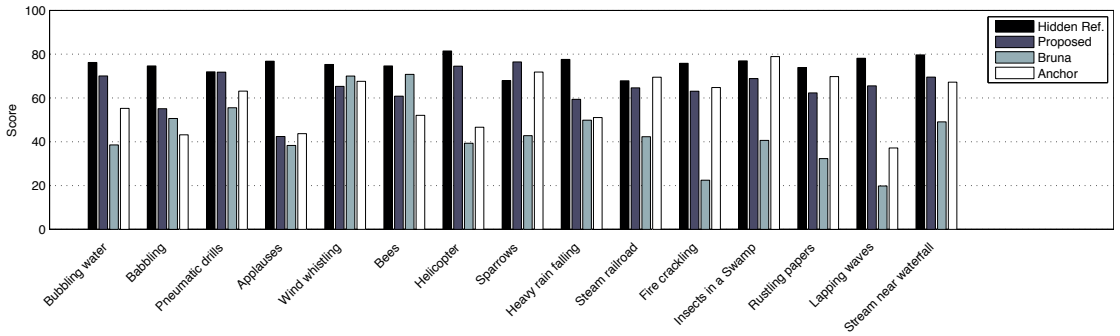
6.3 Discussion

In this chapter, in the profiling section, we have examined the computation efficiency of the proposed algorithm. The SNR measurement shows the accuracy to the optimization process. The first perceptual experiment gives a clue about the relevant length of spectro-temporal information for the auditory system, suggesting that an appropriate length is at least as large as $\pm 819.2ms$.

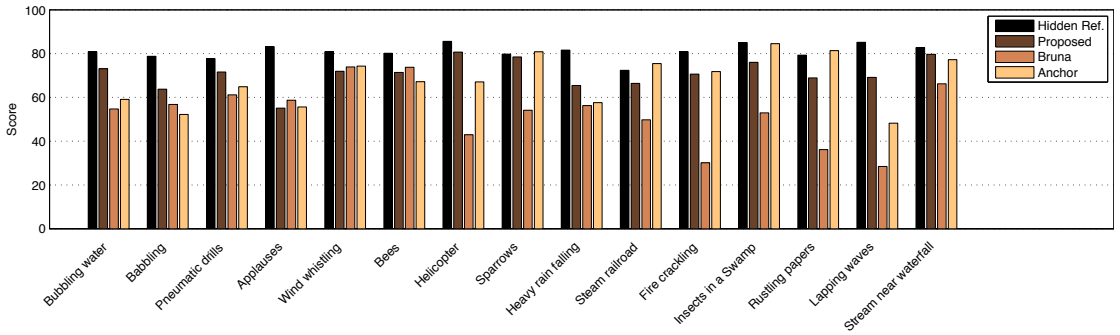
The experiment 2a shows that a representation with small parameter set can still be potential in sound texture processing. From experiment 2b one can know that, with a proper setup, the linear frequency scale signal representation will also be able to compete with the perceptual based signal representation in terms of perceptual quality and computation.

There are several interesting phenomenons. According to the results from Fig. 6.4(d), it can be found that using the headphone can significantly enhance the sensitivity of quality/alikeness to sound textures. From all the results, one can found that expert listeners have a higher chance to identify the hidden references. Besides that, the expert and non-expert listeners have about the same perceptual preferences. The spectrograms of the synthesized sound textures and some examples of prolonged sound textures can be found in Fig. 6.6 and Fig. 6.7.

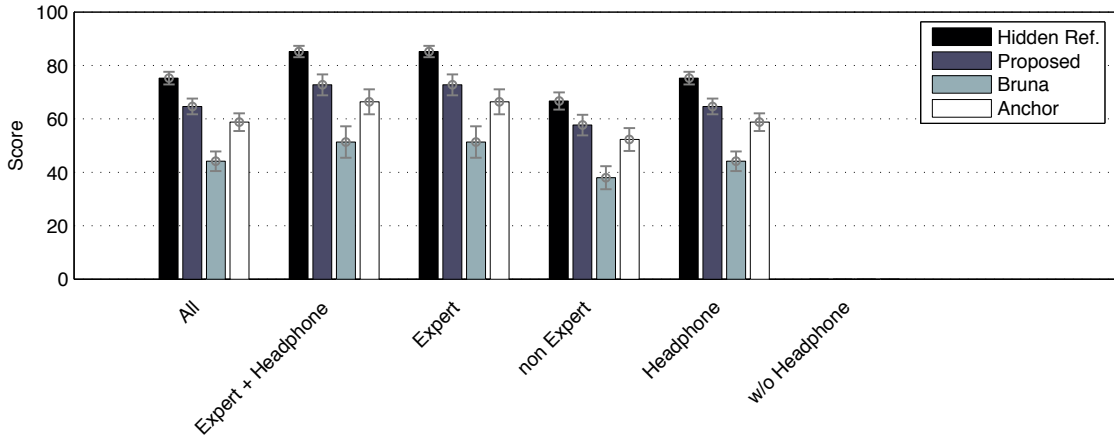
In the perceptual evaluation, the perceptual quality of wide-band events such as clicking of fire and the burst of bubble is not sufficient. These sudden events require high phase synchronization across frequency bands. It is possible that human recognizes these events with something other than time-averaged statistics. If improvement is desired for the quality of these events, post-processing may be required. Besides, the proposed algorithm cannot reproduce the slow wave events in the *Lapping Waves* texture. However, due to that human can recognize these events, using time-averaged statistics may not be sufficient for the slow waves as well. McDermott's work(McDermott and Simoncelli, 2011a) also mentioned that, human recognizes sound events in the texture with something other than statistics. This might be the similar case as the difficulty encountered in the visual texture(Portilla and Simoncelli, 2000).



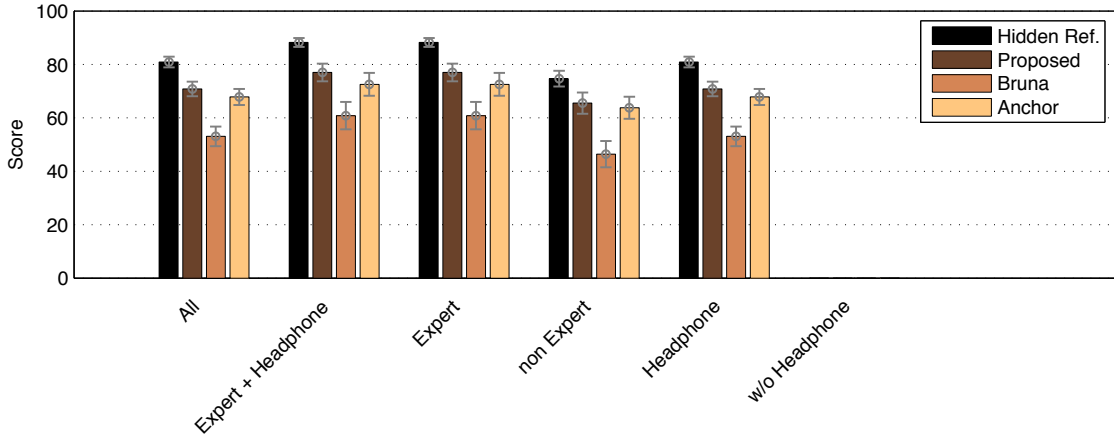
(A) *The quality score of individual files*



(B) *The alikeness score of individual files*

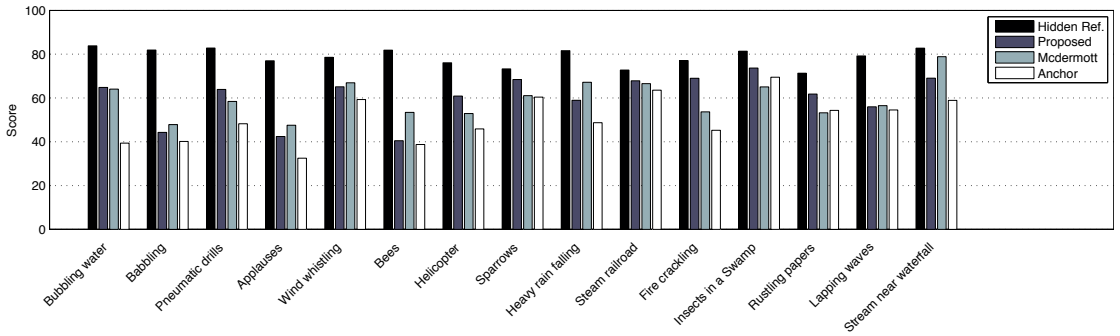


(C) *The quality score averaged over all files(errorbar: 95% confidence interval)*

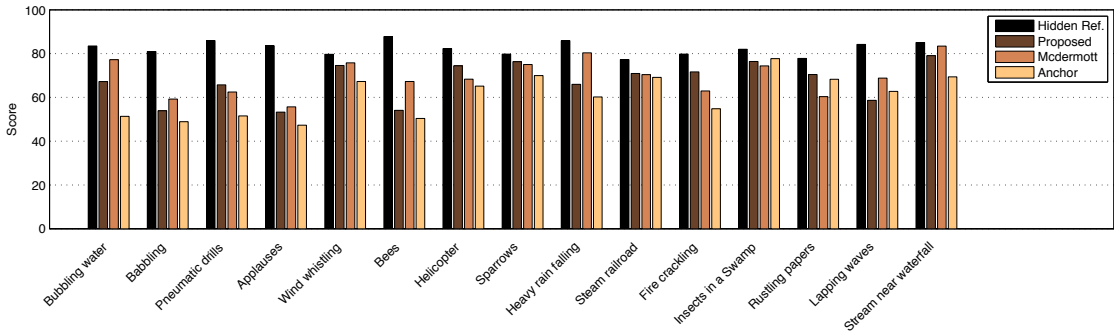


(D) *The alikeness score averaged over all files(errorbar: 95% confidence interval)*

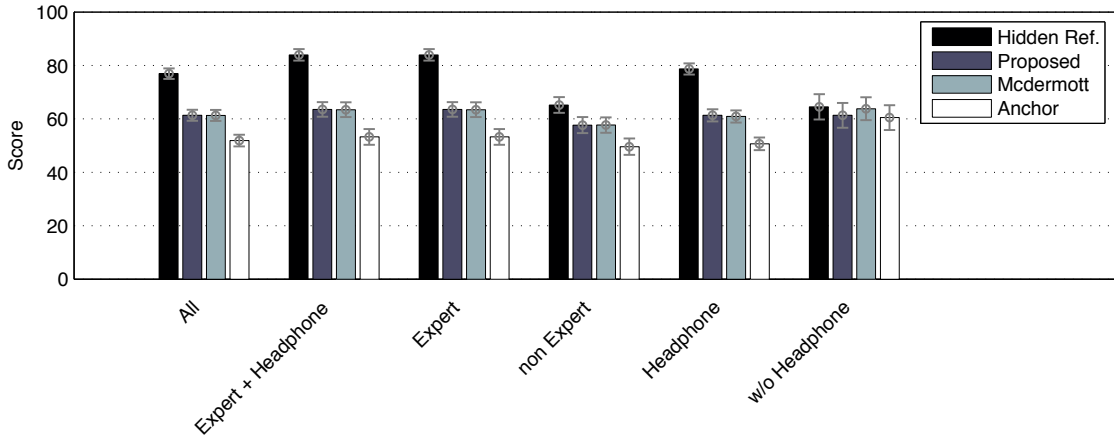
FIGURE 6.4: *The result of experiment 2a .*



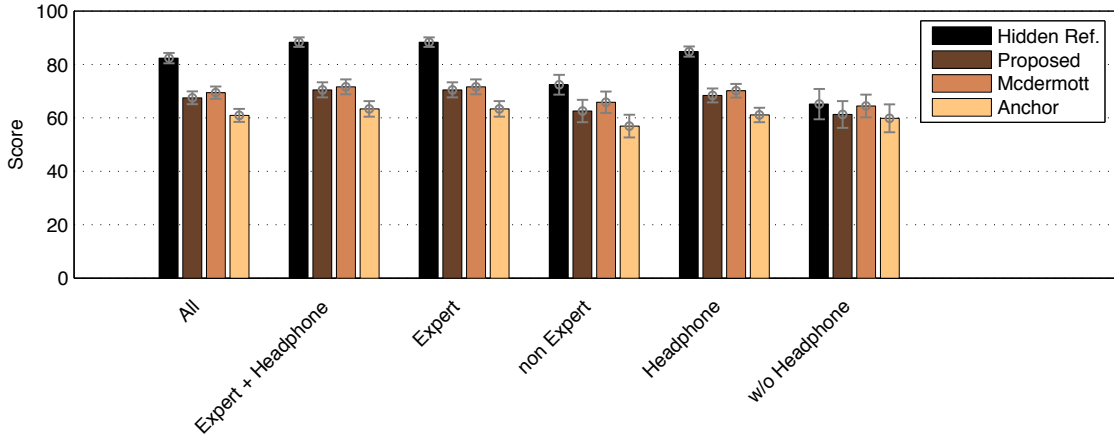
(A) *The quality score of individual files*



(B) *The alikeness score of individual files*

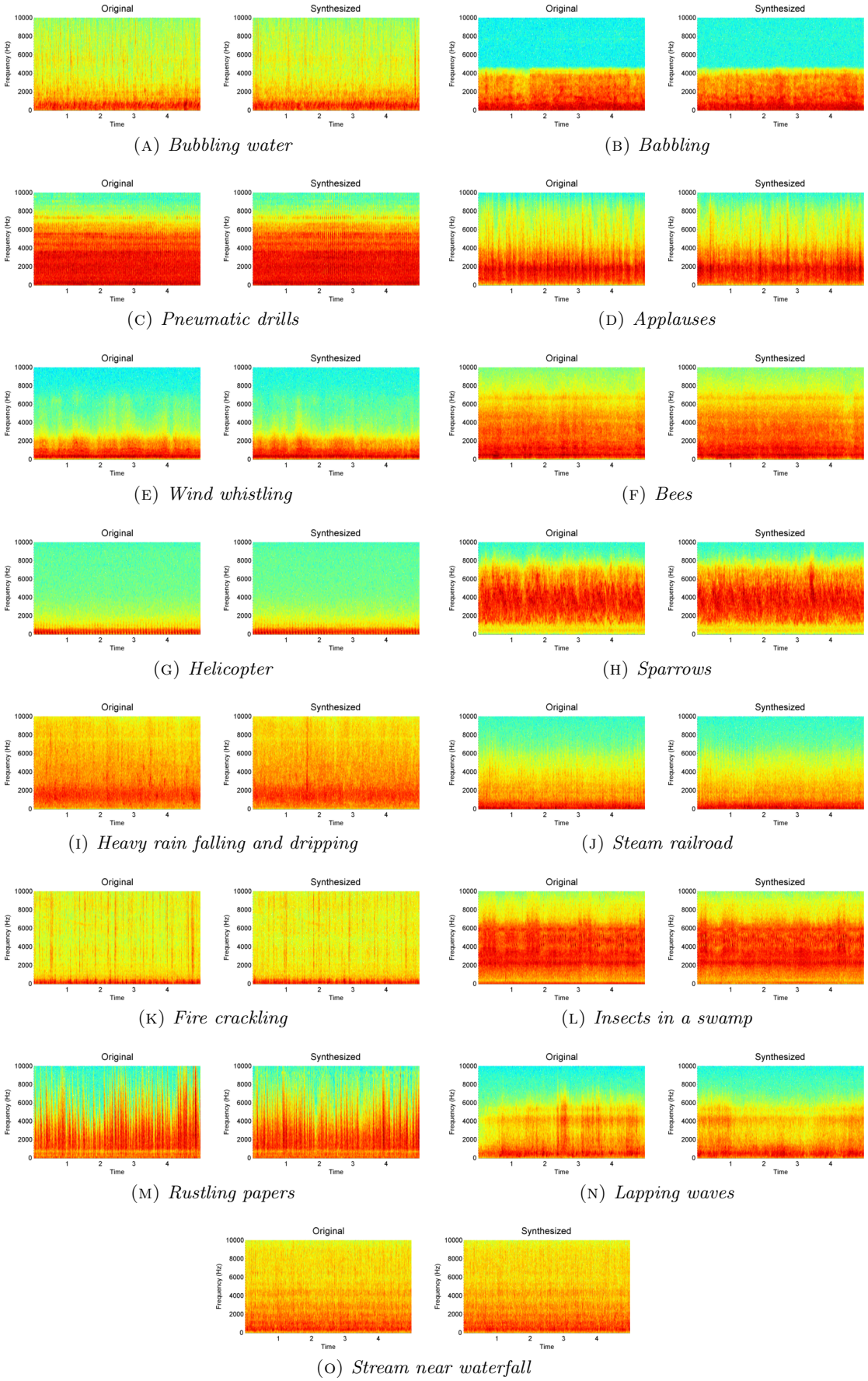


(C) *The quality score averaged over all files(errorbar: 95% confidence interval)*



(D) *The alikeness score averaged over all files(errorbar: 95% confidence interval)*

FIGURE 6.5: *The result of experiment 2b .*



(A) *Bubbling water* (B) *Babbling*

(C) *Pneumatic drills* (D) *Applauses*

(E) *Wind whistling* (F) *Bees*

(G) *Helicopter* (H) *Sparrows*

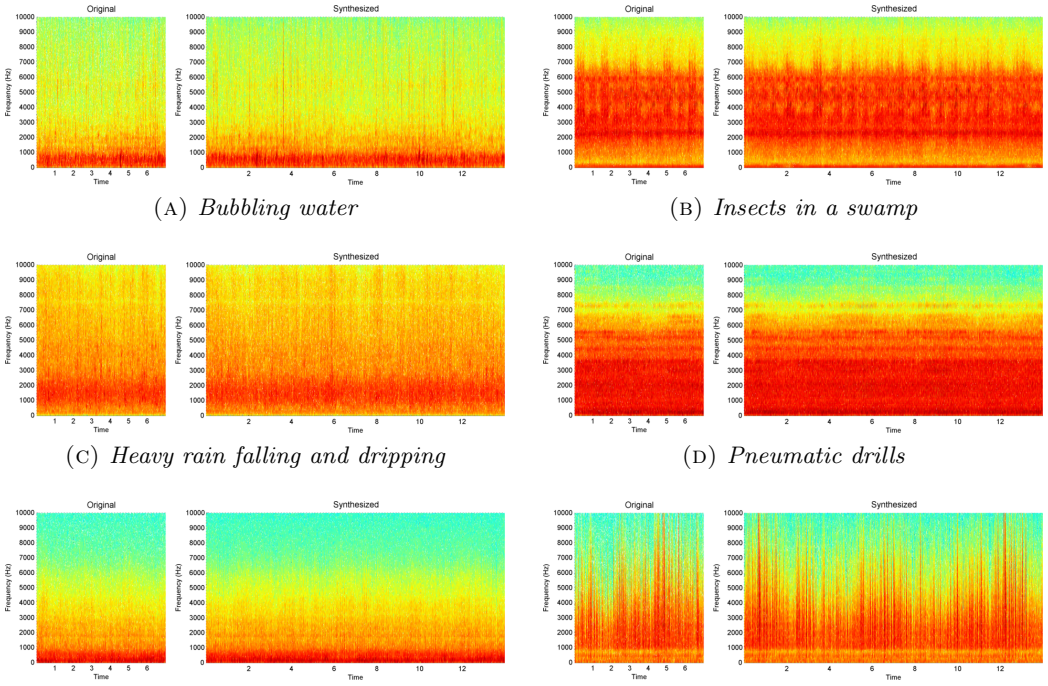
(I) *Heavy rain falling and dripping* (J) *Steam railroad*

(K) *Fire crackling* (L) *Insects in a swamp*

(M) *Rustling papers* (N) *Lapping waves*

(O) *Stream near waterfall*

FIGURE 6.6: *The spectrograms of original and synthetic textures.*



(A) *Bubbling water*

(B) *Insects in a swamp*

(C) *Heavy rain falling and dripping*

(D) *Pneumatic drills*

(E) *Steam railroad*

(F) *Rustling papers*

FIGURE 6.7: *Some prolonged examples. The spectrograms of original(left) and generated(right) textures.*

Chapter 7

Conclusion & Perspectives

7.1 Conclusion

This thesis proposed an analysis-synthesis framework that is suitable for most of sound textures. The framework includes a statistical description that characterizes a sound texture with its perceptually important statistics and an algorithm to generate new sound texture samples from the statistical description. The statistical description was not bounded on any specific transform or time-frequency representation, as long as the underlying representation fits some conditions like invertibility and data grid regularity. The perceptual experiment also shows that, with proper characteristic preserved, one can use traditional linear frequency scale representations to generate samples efficiently that rivals the samples generated from perceptual frequency scaled representation in terms of perceptual quality or similarity. In the subjective evaluation, the perceptual test found that the relevant length of spectro-temporal correlation to the auditory system is at least $\pm 819.2ms$.

Regarding the adaptation of statistics of a signal, we proposed a method to impose every correlation function in a time-frequency representation to another signal at once. This method also helps in understanding how the statistics and the signals are related. If the autocorrelation is fixed, then the cross-correlation function can be controlled by adapting the phase differences between two spectra. We also discovered the relation between the moments and the spectrum of signal. The moments can be represented with the correlation functions of the spectral coefficients. This helps to establish a moment imposition algorithm that preserves correlation functions when imposing statistical moments. In order to allow more variety of the generated sound texture, we proposed the partial cross-correlation function imposition. It imposes the terms of cross-correlation functions selected by a weighting vector. This imposition of cross-correlation preserves

autocorrelation function in the process. With these tools to impose the statistics, we can have an iterative algorithm that imposes autocorrelation, cross-correlation and statistical moments sequentially. The stage-by-stage statistic imposition will preserve all the statistics in the previous stage, which can reduce the computation intensity and reduce the computation time for the optimization process. Therefore, the proposed algorithm generates new samples that have the same perceptual quality as McDermott's algorithm with a much shorter time.

In the article, we also discussed about the difficulties encountered for the proposed algorithm. This includes the necessity of phase reconstruction algorithm, the complex non-differentiability of objective function and the difficulty to preserve the non-negativity of the generated sub-band signal. It is also mentioned in the evaluation section that the perceptual quality of short-time sudden events is not sufficient. However, this is related to the question that, if there are audible events in sound textures, can we deal with the events with only the statistics? This would require further investigation.

According to the evaluation, the proposed analysis-synthesis framework rivals the state-of-the-art algorithms including McDermott's work in terms of the quality of generated sound textures. It is also faster than McDermott's algorithm and is possible to be faster if proper optimization is done.

7.2 Perspectives

There are several ways to improve the proposed analysis-synthesis framework. Some of them related to the optimization, some of them related to the controllability.

About the optimization, it is possible to speed-up the partial cross-correlation imposition if one can efficiently generate a spectrum of sub-band signal that has similar cross-correlation functions as the target in the statistical description. This would greatly reduce the amount of iterations required to reach a local minimum thus accelerate the whole synthesis process. Another possible improvement is to consider the effect of analysis window when generating the sub-band signals. This will make the phase reconstruction a part of the algorithm thus remove the requirement of phase reconstruction and reduce the unresolvable inconsistencies in the time-frequency domain. The third possible improvement is to combine the partial cross-correlation imposition and the spectral moment imposition into one optimization process. This is beneficial because in the current state, the algorithm prioritizes cross-correlation function over the moments. The weighting of moments may be too small. If the moments are imposed

along with the cross-correlation functions, another weighting vector can be used to control the weighting between these two. This may be achieved by studying the Hessian matrix of the objective function and put proper weighting parameters in different parts of the objective function. A further improvement would be to be able to impose partial autocorrelation functions. The partial imposition of autocorrelation function changes only the magnitude spectra of sub-band signals and can be combined with the spectral moment imposition. This partial auto-correlation imposition can be combined with the improved partial cross-correlation imposition to form a new optimization process, which updates alternatively between these two.

The next improvement is related to the statistical description. As it is stated in chapter 3 that the cross-correlation component, $\mathcal{C}_{\mathcal{X},\mathcal{X}}$, in the statistical description contains redundant information. This means we cannot assign arbitrary values to cross-correlation functions without inducing inconsistencies. If we can find a more efficient representation of $\mathcal{C}_{\mathcal{X},\mathcal{X}}$, it will not only allows the arbitrary value assignment of the cross-correlation but also reduces the redundant information in the statistical description.

About the controllability, since many applications need to synchronize visual events with sound events. It would broaden the utility of the proposed analysis-synthesis framework if we can control the appearance of the events to some extends. Moreover, in the present state, if we would like to alter the result of synthesis, we can only change the statistics directly in the statistical description. It will be much useful if a semantic control can be developed, which links some high level features of a sound to a specific pattern of the statistical description.

There are many potential applications for the proposed analysis-synthesis framework. The most straightforward application is to use the framework to serve as a generator of sound scenes. It is possible to use the proposed framework to interpolate between similar textures and possibly with some gradual changes to achieve texture morphing. Another useful application is to use the proposed algorithm to deal with non-sinusoidal parts in human speech and some instruments. Further application will be using the framework to explore unseen sound textures, these unseen sounds can be used in art compositions or media like games, films or virtual reality.

Appendix A

Raw Moments in Terms of Spectral Correlation Functions

This appendix describes how to calculate moments from the spectrum of a signal and shows the derivatives respect to the spectral phases. This is important if someone wants to impose moments by alternating the spectral phases of a sub-band signal such as the spectral domain imposition in section 4.2.2.

Since the first two moments are already preserved by imposing the autocorrelation, we would begin to investigate how to calculate skewness and kurtosis in the spectra of sub-band signals. In order to figure this out, some simplifications should help. The skewness and kurtosis can be decomposed in these forms:

$$\eta(x) = \frac{\mathrm{E}[x^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3} \tag{A.1}$$

$$\kappa(x) = \frac{\mathrm{E}[x^4] - 4\mu\,\mathrm{E}[x^3] + 3\mu^4 + 6\mu^2\sigma^2}{\sigma^4} \tag{A.2}$$

If the signal is standardized($\mu = 0, \sigma = 1$), all the terms besides the first will disappear. The skewness and kurtosis thus become equal to the third and fourth raw moments. Then we can derive how to calculate raw moments in the spectrum of a signal. This is rather easy for the mean and the second raw moment, as shown below:

$$\mathrm{E}[x] = \frac{1}{N_x}\hat{x}(0) \tag{A.3}$$

$$\mathrm{E}[x^2] = \frac{1}{N_x^2}\sum_{k=0}^{N_k-1}\overline{\hat{x}}(k)\hat{x}(k) \quad \text{(Parseval's theorem)} \tag{A.4}$$

N_x is the length of x, N_k is the length of spectrum. According to the convolution theorem and the associativity of convolution, we know that:

$$\mathcal{F}\{x^n \cdot x^m\} = \frac{1}{N_x}\mathcal{F}\{x^n\} * \mathcal{F}\{x^m\} \quad \text{(convolution theorem)} \tag{A.5}$$

$$(\mathcal{F}\{x^n\} * \mathcal{F}\{x^m\}) * \mathcal{F}\{x^o\} = \mathcal{F}\{x^n\} * (\mathcal{F}\{x^m\} * \mathcal{F}\{x^o\}) \quad \text{(associativity)} \tag{A.6}$$

Since x is real, $\mathcal{F}\{x^n\}$ must be Hermitian. In a similar fashion to (A.4), it can be shown that:

$$\begin{aligned}\mathrm{E}[x^{n+m}] &= \frac{1}{N_x}(\frac{1}{N_x}\mathcal{F}\{x^n\} * \mathcal{F}\{x^m\})_0 \\ &= \frac{1}{N_x}\mathrm{E}[\overline{\mathcal{F}}\{x^n\} \cdot \mathcal{F}\{x^m\}] \\ &= \frac{1}{N_x^2}\sum_{k=0}^{N_k-1}\overline{\mathcal{F}}\{x^n\}_k \cdot \mathcal{F}\{x^m\}_k\end{aligned} \tag{A.7}$$

The last piece we need is again the convolution theorem:

$$\mathcal{F}\{x^2\}_k = (\mathcal{F}\{x\} * \mathcal{F}\{x\})_k = \frac{1}{N_x}A_{\hat{x}}(k) \tag{A.8}$$

In (A.8), we see the appearance of the autocorrelation of the spectrum of the signal. In order to make it short, we will refer this as **spectral autocorrelation**. Combining

(A.7) and (A.8), now we can derive the equations for the third and fourth raw moments:

$$\begin{aligned}
\mathrm{E}[x^3] &= \mathrm{E}[x^2 \cdot x] \\
&= \frac{1}{N_x^2} \sum_{k=0}^{N_k-1} \overline{\mathcal{F}}\{x^2\}_k \cdot \mathcal{F}\{x\}_k \\
&= \frac{1}{N_x^3} \sum_{k=0}^{N_k-1} \overline{A}_{\hat{x}}(k)\hat{x}(k) \qquad \text{(A.9)} \\
\mathrm{E}[x^4] &= \mathrm{E}[x^2 \cdot x^2] \\
&= \frac{1}{N_x^2} \sum_{k=0}^{N_k-1} \overline{\mathcal{F}}\{x^2\}_k \cdot \mathcal{F}\{x^2\}_k \\
&= \frac{1}{N_x^4} \sum_{k=0}^{N_k-1} \overline{A}_{\hat{x}}(k)A_{\hat{x}}(k) \qquad \text{(A.10)}
\end{aligned}$$

It is interesting that the third raw moment is proportional to the sum of product of the conjugated spectral autocorrelation and the spectrum itself; the fourth raw moment is the sum of product of the spectral autocorrelation and its conjugated version. In fact, the similar fashion can be shown for higher order of raw moments($n > 4$):

$$A_x^2 = A_{A_x}, \quad A_x^3 = A_{A_{A_x}}, \quad A_x^{(n)} = \overbrace{A_{A_{A\cdots A_x}}}^{n} \qquad \text{(A.11)}$$

$$\mathrm{E}[x^n] = \frac{1}{N_x^n} \sum_{k=0}^{N_k-1} \overline{A}_{\hat{x}}^{(\frac{n-1}{2})}(k)\hat{x}(k), \quad n \bmod 2 \equiv 1 \qquad \text{(A.12)}$$

$$\mathrm{E}[x^n] = \frac{1}{N_x^n} \sum_{k=0}^{N_k-1} \overline{A}_{\hat{x}}^{(\frac{n-2}{2})}(k)A_{\hat{x}}(k), \quad n \bmod 2 \equiv 0 \qquad \text{(A.13)}$$

In (A.11), the notation A_{A_x} means the *autocorrelation of autocorrelation of* x, which also written as A_x^2. Therefore, A_x^3 is the autocorrelation of A_x^2. The rest of (A.11) may be deduced by analogy.

Besides evaluating the moment itself, if we would like to use (A.9) and (A.10) in an optimization process, it will be necessary to know the derivatives respect to the change of phases of $\hat{x}$. To clarify, $\hat{x}$ can be written as $\hat{x}(k) = |\hat{x}_k|e^{j\theta_k}$. We would like to find the derivatives respect to θ. Remembering x is real-valued, so $\hat{x}$ and $A_{\hat{x}}$ must be Hermitian. Therefore, it must keep in mind that the conjugated components must be dealt together when deriving the derivatives. A useful fact related to the conjugated components is:

$$ze^{j\theta_\alpha} - \overline{z}e^{-j\theta_\alpha} = 2j\Im\{ze^{j\theta_\alpha}\} \qquad \text{(A.14)}$$

We will begin from the third raw moment. Recall (A.9), the derivative of $\overline{A}_{\hat{x}}(k)$ respect to $x(\alpha)$ is:

$$\frac{\partial}{\partial \hat{x}(\alpha)} \overline{A}_{\hat{x}}(k) = \frac{\partial}{\partial x(\alpha)} (\overline{\hat{x}}(\alpha)\hat{x}(\alpha+k) + \overline{\hat{x}}(\alpha-k)\hat{x}(\alpha)) \tag{A.15}$$

Next, deal with $\overline{A}_{\hat{x}}(k)\hat{x}(k)$:

$$\begin{aligned} &\frac{\partial}{\partial \hat{x}(\alpha)} \overline{A}_{\hat{x}}(k)\hat{x}(k) \\ &= \begin{cases} \frac{\partial}{\partial \hat{x}(\alpha)} (\overline{\hat{x}}(\alpha)\hat{x}(\alpha+k) + \overline{\hat{x}}(\alpha-k)\hat{x}(\alpha))\hat{x}(k) & k \neq \alpha \\ \frac{\partial}{\partial \hat{x}(\alpha)} \Big(\overline{A}_{\hat{x}}(\alpha)\hat{x}(\alpha) + (\overline{\hat{x}}(\alpha)\hat{x}(2\alpha) + \overline{\hat{x}}(0)\hat{x}(\alpha))\hat{x}(\alpha) \Big) & k = \alpha \end{cases} \end{aligned} \tag{A.16}$$

Proceed to the main body of (A.9):

$$\frac{\partial}{\partial \hat{x}(\alpha)} \sum_{k=0}^{N_k-1} \overline{A}_{\hat{x}}(k)\hat{x}(k) = \frac{\partial}{\partial \hat{x}(\alpha)} \Big(\overline{A}_{\hat{x}}(\alpha)\hat{x}(\alpha) + \sum_{k=0}^{N_k-1} \hat{x}(k)(\overline{\hat{x}}(\alpha)\hat{x}(\alpha+k) + \overline{\hat{x}}(\alpha-k)\hat{x}(\alpha)) \Big) \tag{A.17}$$

Because of $\hat{x}$ and $A_{\hat{x}}$ are Hermitian, therefore:

$$\overline{\hat{x}}(\alpha) \sum_{k=0}^{N_k-1} \hat{x}(k)\hat{x}(k+\alpha) = \overline{\hat{x}}(\alpha) A_{\hat{x}}(\alpha) \tag{A.18}$$

$$\hat{x}(\alpha) \sum_{k=0}^{N_k-1} \hat{x}(k)\hat{x}(k-\alpha) = \hat{x}(\alpha) \overline{A}_{\hat{x}}(\alpha) \tag{A.19}$$

Thus, the right hand side of (A.17) becomes:

$$\frac{\partial}{\partial \hat{x}(\alpha)} (2\overline{A}_{\hat{x}}(\alpha)\hat{x}(\alpha) + \overline{\hat{x}}(\alpha) A_{\hat{x}}(\alpha)) \tag{A.20}$$

Since we are actually differentiating respect to θ_α, we need:

$$\frac{\partial \hat{x}(\alpha)}{\partial \theta_\alpha} = j\hat{x}(\alpha), \quad \frac{\partial \overline{\hat{x}}(\alpha)}{\partial \theta_\alpha} = -j\hat{x}(\alpha) \tag{A.21}$$

Now, put everything including the conjugated components together. The conjugated component always moves in the opposite direction along the circle of amplitude in the

complex plane when the angle is changed:

$$\begin{aligned}&\frac{\partial}{\partial\theta_\alpha}\sum_{k=0}^{N_k-1}\overline{A}_{\hat{x}}(k)\hat{x}(k)\\&=j(2\overline{A}_{\hat{x}}(\alpha)\hat{x}(\alpha)-\overline{\hat{x}}(\alpha)A_{\hat{x}}(\alpha))-j(2A_{\hat{x}}(\alpha)\overline{\hat{x}}(\alpha)-\hat{x}(\alpha)\overline{A}_{\hat{x}}(\alpha))\end{aligned}\tag{A.22}$$

Apply (A.14), and putting back the scalar multiplier in (A.9), we have:

$$\frac{\partial}{\partial\theta_\alpha}\frac{1}{N_x^3}\sum_{k=0}^{N_k-1}\overline{A}_{\hat{x}}(k)\hat{x}(k)=\frac{-6}{N_x^3}\Im\{\overline{A}_{\hat{x}}(\alpha)\hat{x}(\alpha)\}\tag{A.23}$$

After dealing with the third raw moment, we will use similar steps to proceed with the fourth moment(apply (A.15)).

$$\begin{aligned}\frac{\partial}{\partial\theta_\alpha}\sum_{k=0}^{N_k-1}\overline{A}_{\hat{x}}(k)A_{\hat{x}}(k)=\frac{\partial}{\partial\theta_\alpha}\sum_{k=0}^{N_k-1}\Big(&\overline{A}_{\hat{x}}(k)(\hat{x}(\alpha)\overline{\hat{x}}(\alpha+k)+\hat{x}(\alpha-k)\overline{\hat{x}}(\alpha))\\&+A_{\hat{x}}(k)(\overline{\hat{x}}(\alpha)\hat{x}(\alpha+k)+\overline{\hat{x}}(\alpha-k)\hat{x}(\alpha))\Big)\end{aligned}\tag{A.24}$$

Concatenate and rewrite the terms in the form of cross-correlation, the right hand side becomes:

$$\frac{\partial}{\partial\theta_\alpha}2\hat{x}(\alpha)C_{\overline{A}_{\hat{x}},\hat{x}}(\alpha)+2\overline{\hat{x}}(\alpha)C_{\overline{A}_{\hat{x}},\hat{x}}(-\alpha)\tag{A.25}$$

This time, the conjugated component is exactly the same. Therefore, we have:

$$\frac{\partial}{\partial\theta_\alpha}\sum_{k=0}^{N_k-1}\overline{A}_{\hat{x}}(k)A_{\hat{x}}(k)=4j\hat{x}(\alpha)C_{\overline{A}_{\hat{x}},\hat{x}}(\alpha)-4j\overline{\hat{x}}(\alpha)\overline{C}_{\overline{A}_{\hat{x}},\hat{x}}(\alpha)\tag{A.26}$$

Again, apply (A.14), now we obtained the gradient of the fourth raw moment:

$$\frac{\partial}{\partial\theta_\alpha}\frac{1}{N_x^4}\sum_{k=0}^{N_k-1}\overline{A}_{\hat{x}}(k)A_{\hat{x}}(k)=\frac{-8}{N_x^4}\Im\{\overline{C}_{A_{\hat{x}},\hat{x}}(\alpha)\hat{x}(\alpha)\}\tag{A.27}$$

With (A.23) and (A.27) in hand, we can now apply (A.9) and (A.10) in optimization processes. To be mentioned, recalling the assumption we made at the beginning, the signal must be standardized. In other words, the signal must be zero-mean and have unit-variance. In the practical case, since the mean of signal x is zero, $\hat{x}(0)$ is also zero, thus will be skipped in the optimization process. If the standardization is undesired, the

gradient must be changed according to (A.1) and (A.2). We also know that:

$$\frac{\partial}{\partial \theta_\alpha}\hat{x}(0) = 0 \quad \alpha \neq 0 \tag{A.28}$$

$$\frac{\partial}{\partial \theta_\alpha}\frac{1}{N_x^2}\sum_{k=0}^{N_k-1}\overline{\hat{x}}(k)\hat{x}(k) = \frac{-4}{N_x^2}\Im\{\overline{\hat{x}}(\alpha)\hat{x}(\alpha)\} \tag{A.29}$$

Thus, if necessary, the gradient of the moment can be derived by combining the gradient of each individual term.

Appendix B

The Complex Differentiability of the partial correlation imposition

In this appendix, we would like to discuss the complex differentiability of the objective function of partial correlation imposition. This objective function attempts to adapt both autocorrelation function and cross-correlation function in the spectral domain of the sub-band signals. Adapting both types of correlation functions at the same time in the spectral domain means that we need a gradient respect to a complex-valued vector. Unfortunately, we are going to show that this objective function is not complex-differentiable, in other words, a gradient, which can guide an optimization process, does not exist for this objective function.

With no loss of generality, we can consider a simplified version of the partial correlation function imposition. There are two real valued signals, x and y. We would like to change $\hat{x}$ in order to impose an autocorrelation function $\mathcal{A}$ on x and adjust the cross-correlation function between x,y such that $C_{x,y} = \mathcal{C}$. The objective function will be looked like this:

$$\begin{aligned} f(\hat{x}, y, \mathcal{A}, \mathcal{C}) &= \sum_{\tau=-\infty}^{\infty} |A_x(\tau) - \mathcal{A}(\tau)|^2 + \sum_{\tau=-\infty}^{\infty} |C_{x,y}(\tau) - \mathcal{C}(\tau)|^2 \\ &= \sum_{k=0}^{N_k-1} |\hat{A}_x(k) - \hat{\mathcal{A}}(k)|^2 + \sum_{k=0}^{N_k-1} |(\hat{x} \circ \overline{\hat{y}})_k - (\hat{\mathcal{C}})_k|^2 \end{aligned} \tag{B.1}$$

The derivative of (B.1) respect to $\hat{x}(\alpha)$ is (conjugate component included):

$$\begin{aligned}\partial f(\hat{x}, y, \mathcal{A}, \mathcal{C}) &= 2(\hat{A}_x(\alpha) - \hat{\mathcal{A}}(\alpha))\partial(\overline{\hat{x}}(\alpha)\hat{x}(\alpha)) + 2(\hat{x}(\alpha)\overline{\hat{y}}(\alpha) - \hat{\mathcal{C}}(\alpha))\overline{\hat{y}}(\alpha)\partial\hat{x}(\alpha) \\ &+ 2(\hat{A}_x(\alpha) - \hat{\mathcal{A}}(\alpha))\partial(\overline{\hat{x}}(\alpha)\hat{x}(\alpha)) + 2(\overline{\hat{x}}(\alpha)\hat{y}(\alpha) - \overline{\hat{\mathcal{C}}}(\alpha))\hat{y}(\alpha)\partial\overline{\hat{x}}(\alpha) \quad \text{(B.2)}\end{aligned}$$

If a complex function is differentiable, it must follow the Cauchy-Riemann equation(Markushevich, 2005):

$$\begin{aligned}&f(x) = u(x) + i \cdot v(x), \quad x = a + ib \\ &\frac{\partial u}{\partial a} = \frac{\partial v}{\partial b}, \quad \frac{\partial u}{\partial b} = -\frac{\partial v}{\partial a} \quad \text{or} \quad i\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b}\end{aligned} \tag{B.3}$$

The meaning of Cauchy-Riemann equation is that, the gradient of a point must be the same regardless which direction was used to approach the point. We can examine whether (B.2) fits the Cauchy-Riemann equation:

$$\begin{aligned}\hat{x}(\alpha) &\equiv a + ib \\ \Delta A(\alpha) &= \hat{A}_x(\alpha) - \hat{\mathcal{A}}(\alpha), \quad \Delta A(\alpha) \in \mathbb{R} \\ \Delta C(\alpha) &= \hat{x}(\alpha)\overline{\hat{y}}(\alpha) - \hat{\mathcal{C}}(\alpha), \quad \Delta C(\alpha) \in \mathbb{C} \\ \partial f(\hat{x}, y, \mathcal{A}, \mathcal{C}) &= 4\Delta A(\alpha)\partial(a^2 + b^2) + 2\Delta C(\alpha)\overline{\hat{y}}(\alpha)\partial(a + ib) + 2\overline{\Delta C}(\alpha)\hat{y}(\alpha)\partial(a - ib)\end{aligned} \tag{B.4}$$

$$i\frac{\partial f}{\partial a} = 8ia\Delta A(\alpha) + 2i\Re\{\Delta C(\alpha)\overline{\hat{y}}(\alpha)\} \tag{B.5}$$

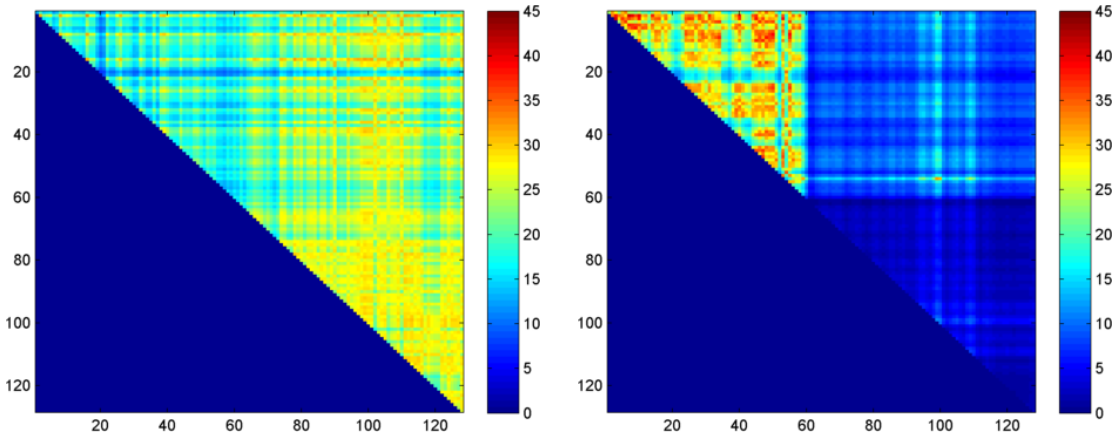
$$\frac{\partial f}{\partial b} = 8b\Delta A(\alpha) + 2i\Im\{\Delta C(\alpha)\overline{\hat{y}}(\alpha)\} \tag{B.6}$$

Unfortunately, (B.5) and (B.6) will only equal to each other if $a = ib$ and $\Re\{\Delta C(\alpha)\overline{\hat{y}}(\alpha)\} = \Im\{\Delta C(\alpha)\overline{\hat{y}}(\alpha)\}$. However, the condition cannot hold in general. Therefore, in the most of cases, (B.1) is non-complex differentiable. It means that there is no reliable directive in the error plane of (B.1). Therefore, we cannot use existing optimization techniques to adapt correlation functions in the spectral domain of sub-band signals. This is a limit of adapting correlation functions in the spectral domain of sub-band signal.

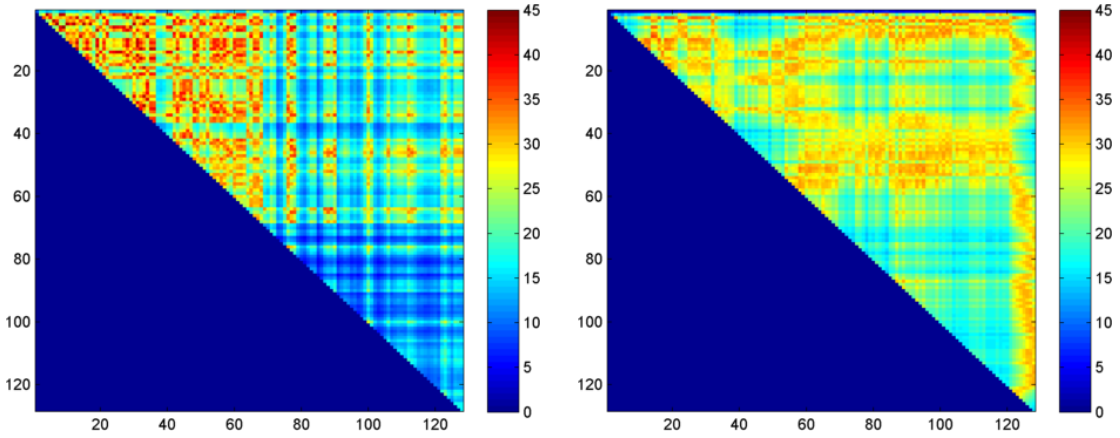
Appendix C

The SNR(Signal-toNoise Ratio) of Correlation Functions for the Sound Texture Samples

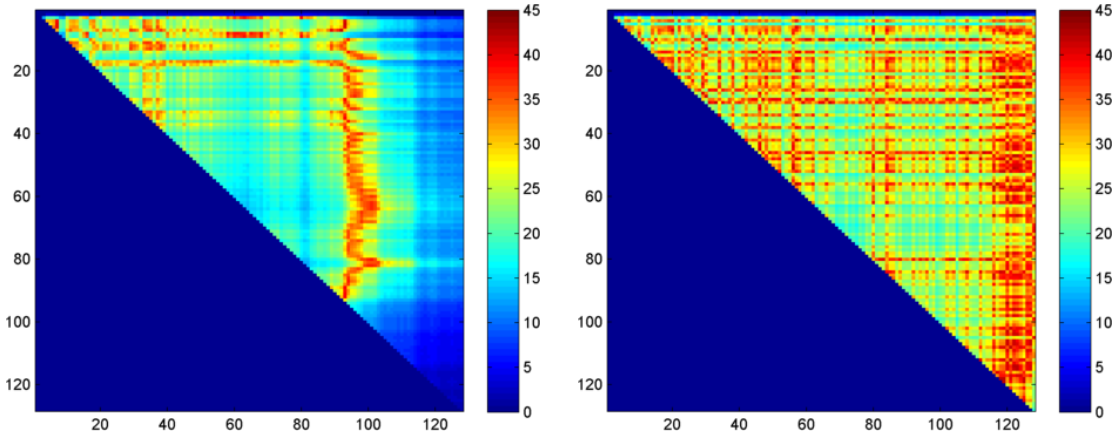
This appendix shows the SNR of correlation functions in dB scale. The x and y axis indicate the index of the sub-band signal. For example, the color of point 20,24 indicates the SNR of the cross-correlation function between 20th and 24th sub-bands. The color represents the SNR in dB scale.



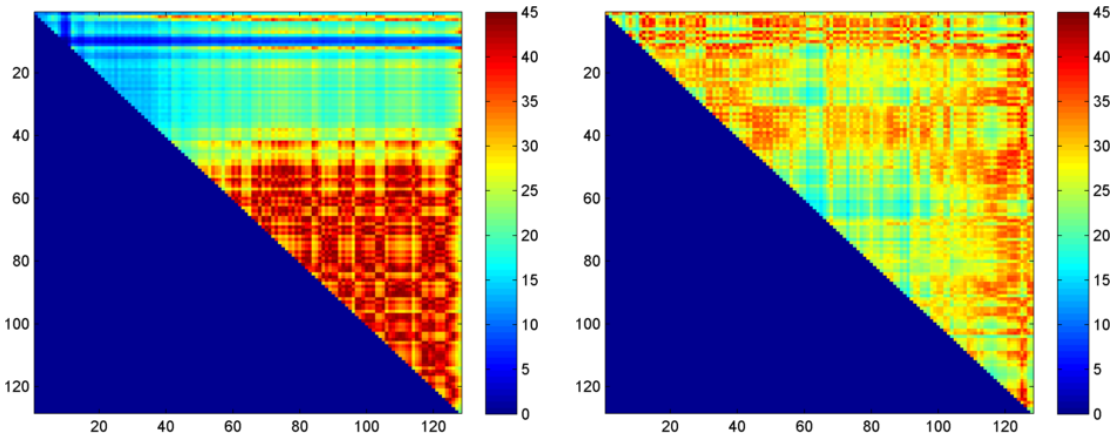
(A) *Bubbling water* (B) *Babbling*



(C) *Pneumatic drills* (D) *Applauses*

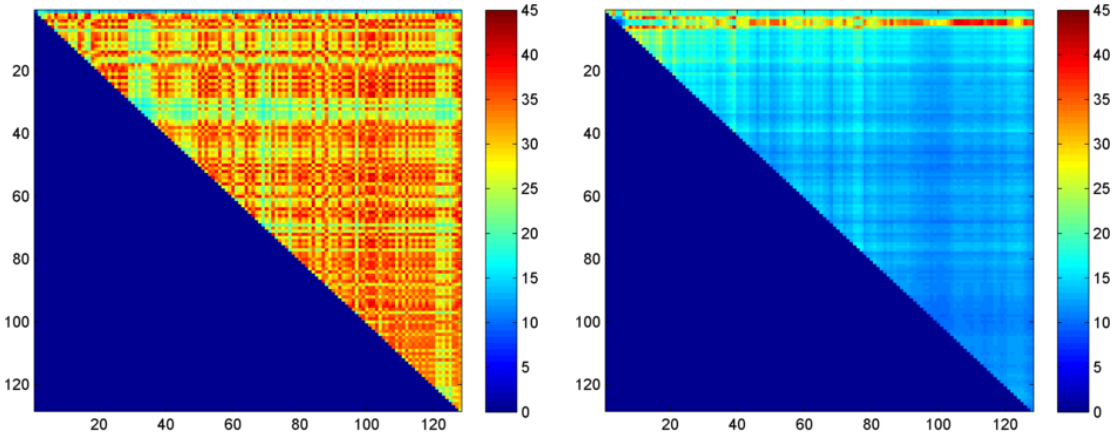


(E) *Wind whistling* (F) *Bees*



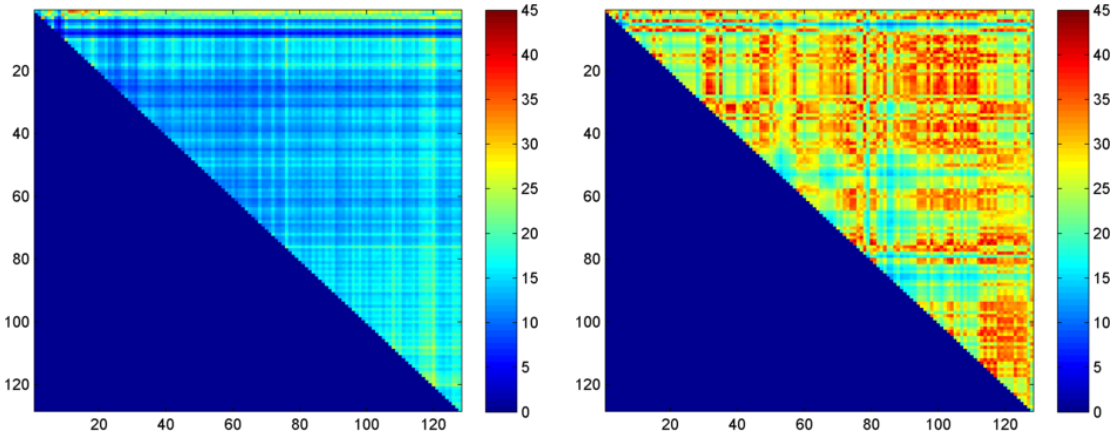
(G) *Helicopter*

(H) *Sparrows*



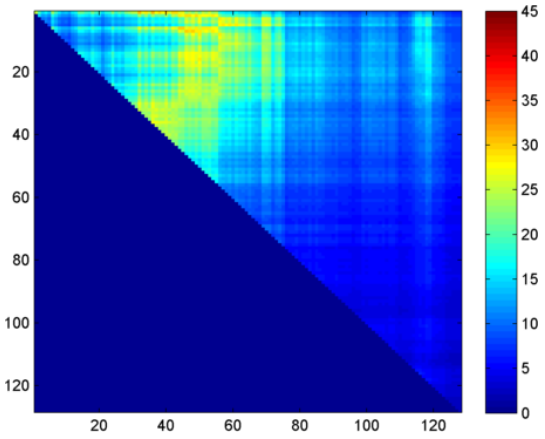
(I) *Heavy rain falling and dripping*

(J) *Steam railroad*

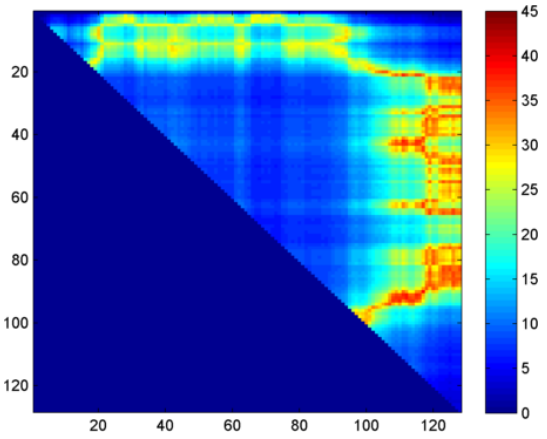


(K) *Fire crackling*

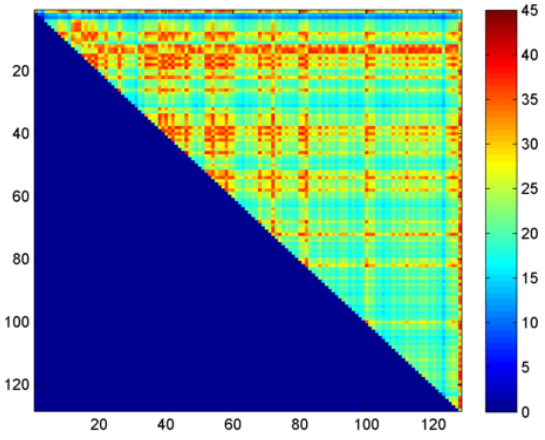
(L) *Insects in a swamp*



(M) *Rustling papers*



(N) *Lapping waves*



(O) *Stream near waterfall*

FIGURE C.-2: *The SNR of the correlation functions of the tested sound texture samples. Axes are the frequency bin indices, with 128 bins each side. The diagonal corresponds to autocorrelation functions; the rest of the upper triangle corresponds to the cross-correlation functions.*

Bibliography

Allen, J. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(3): 235–238, Jun 1977. ISSN 0096-3518. doi: 10.1109/TASSP.1977.1162950.

Apel, T. Sinusoidality analysis and noise synthesis in phase vocoder based time-stretching. In *Proceedings of the Australasian Computer Music Conference*, pages 7–12, 2014.

Athineos, M. and Ellis, D. P. W. Sound texture modelling with linear prediction in both time and frequency domains. In *in Proc. ICASSP*, pages 648–651, 2003.

Attias, H. and Schreiner, C. E. Coding of naturalistic stimuli by auditory midbrain neurons. *Advances in neural information processing systems*, pages 103–109, 1998.

Balazs, P., Dörfler, M., Jaillet, F., Holighaus, N., and Velasco, G. Theory, implementation and applications of nonstationary gabor frames. *Journal of Computational and Applied Mathematics*, 236(6):1481 – 1496, 2011. ISSN 0377-0427. doi: http://dx.doi.org/10.1016/j.cam.2011.09.011. URL http://www.sciencedirect.com/science/article/pii/S0377042711004900.

Bloit, J., Rasamimanana, N., and Bevilacqua, F. Towards morphological sound description using segmental models. In *Proceedings of DAFx*, 2009.

Brown, J. Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America*, 89(1):425–434, January 1991.

Bruna, J. and Mallat, S. Classification with invariant scattering representations. In *The Computing Research Repository (CoRR)*, volume abs/1112.1120, 2011.

Bruna, J. and Mallat, S. Audio texture synthesis with scattering moments. In *The Computing Research Repository (CoRR)*, volume abs/1311.0407, 2013. URL http://dblp.uni-trier.de/db/journals/corr/corr1311.html#BrunaM13.

Chi, T., Ru, P., and Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, 118(2):887–906, 2005. doi: 10.1121/1.1945807. URL http://dx.doi.org/10.1121/1.1945807.

Cohen, L. *Time Frequency Analysis: Theory and Applications*. Prentice Hall PTR, 1995.

Conan, S., Thoret, E., Aramaki, M., Derrien, O., Gondre, C., Kronland-Martinet, R., and Ystad, S. Navigating in a space of synthesized interaction-sounds: Rubbing, scratching and rolling sounds. In *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, York, United Kingdom, September 2013.

Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J Neurophysiol*, 85(3):1220–1234, Mar. 2001. ISSN 0022-3077. URL http://jn.physiology.org/cgi/content/abstract/85/3/1220.

Dolson, M. The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27, 1986.

Dörfler, M., Holighaus, N., Grill, T., and Velasco, G. Constructing an invertible constant-q transform with nonstationary gabor frames. In *6th International Conference on Digital Audio Effects (DAFx)*, pages 344–349, London, United Kingdom, Septembre 2003. URL http://articles.ircam.fr/textes/Roebel03a.

Dubnov, S., Bar-Joseph, Z., El-Yaniv, R., Lischinski, D., and Werman, M. Synthesizing sound textures through wavelet tree learning. *IEEE Comput. Graph. Appl.*, 22(4): 38–48, July 2002. ISSN 0272-1716. doi: 10.1109/MCG.2002.1016697. URL http://dx.doi.org/10.1109/MCG.2002.1016697.

Feichtinger, H. G. and Strohmer, T. *Gabor Analysis and Algorithms*. Birkhaeuser, 1998. ISBN 0817639594.

Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12):2379–2394, 1987.

Fröjd, M. and Horner, A. Sound texture synthesis using an overlap–add/granular synthesis approach. *J. Audio Eng. Soc*, 57(1/2):29–37, 2009. URL http://www.aes.org/e-lib/browse.cfm?elib=14805.

Glasberg, B. and Moore, B. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, Aug. 1990. ISSN 03785955. doi: 10.1016/0378-5955(90)90170-T. URL http://dx.doi.org/10.1016/0378-5955(90)90170-T.

Griffin, D. and Lim, J. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):236–243, 1984.

Hanna, P. and Desainte-catherine, M. Time scale modification of noises using a spectral and statistical model. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on*, volume 6, pages 181–184, April 2003.

Harris, F. J. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

Holighaus, N., Dorfler, M., Velasco, G., and Grill, T. A framework for invertible, real-time constant-q transforms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(4):775–785, April 2013. ISSN 1558-7916. doi: 10.1109/TASL.2012.2234114.

ITU-R. *BS.1534-2 (Method for the subjective assessment of intermediate quality levels of coding systems)*. International Telecommunications Union., July 2014.

Joris, P., Schreiner, C., and Rees, A. Neural processing of amplitude-modulated sounds. *Physiological reviews*, 84(2):541–577, 2004.

Julesz, B. Visual pattern discrimination. *Information Theory, IRE Transactions on*, 8: 84–92, 1962.

Julesz, B., Gilbert, E., and Victor, J. Visual discrimination of textures with identical third-order statistics. *Biological Cybernetics*, 31(3):137–140, 1978. ISSN 0340-1200. doi: 10.1007/BF00336998. URL http://dx.doi.org/10.1007/BF00336998.

Kersten, S. and Purwins, H. Fire texture sound re-synthesis using sparse decomposition and noise modelling. In *Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12)*, York, United Kingdom, September 2012.

Laroche, J. and Dolson, M. New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications. *Journal of the AES*, 47 (11):928–936, 1999a.

Laroche, J. and Dolson, M. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999b.

Le Roux, J., Ono, N., and Sagayama, S. Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. In *Proceedings of the SAPA 2008 ISCA Workshop on Statistical and Perceptual Audition*, pages 23–28, Sept. 2008.

Le Roux, J., Kameoka, H., Ono, N., and Sagayama, S. Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency. In *Proc. 13th International Conference on Digital Audio Effects (DAFx-10)*, pages 397–403, Sept. 2010.

Liao, W.-H., Roebel, A., and Su, A. W. On stretching gaussian noises with the phase vocoder. In *Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12)*, York, United Kingdom, September 2012.

Liao, W.-H., Roebel, A., and Su, A. W. On the modeling of sound textures based on the stft representation. In *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, York, United Kingdom, September 2013.

Liuni, M., Röbel, A., Romito, M., and Rodet, X. An entropy based method for local time-adaptation of the spectrogram. In *Exploring Music Contents*, pages 60–75. Springer, 2011.

Lorenzi, C., Berthommier, F., and Demany, L. Discrimination of amplitude-modulation phase spectrum. *The Journal of the Acoustical Society of America*, 105(5):2987–2990, 1999.

Markushevich, A. I. *Theory of functions of a complex variable*. AMS Chelsea Publishing, 2005.

McDermott, J. H. and Simoncelli, E. P. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5):926–940, Sep 2011a. doi: 10.1016/j.neuron.2011.06.032.

McDermott, J. H. and Simoncelli, E. P. Texture synthesis examples. 2011b. URL http://mcdermottlab.mit.edu/texture_examples/index.html.

McDermott, J. H., Oxenham, A. J., and Simoncelli, E. P. Sound texture synthesis via filter statistics. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-09)*, pages 297–300, New Paltz, NY, Oct 18-21 2009. IEEE Signal Processing Society. doi: 10.1109/ASPAA.2009.5346467.

McDermott, J. H., Schemitsch, M., and Simoncelli, E. P. Summary statistics in auditory perception. *Nature Neuroscience*, 16(4):493–498, April 2013. doi: 10.1038/nn.3347.

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of Neurophysiology*, 102(6):3329–3339, Dec. 2009. ISSN 1522-1598. doi: 10.1152/jn.91128.2008. URL http://dx.doi.org/10.1152/jn.91128.2008.

Misra, A., Cook, P. R., and Wang, G. A new paradigm for sound design. In *Proceedings of the International Conference on Digital Audio Effects (DAFx-06)*, pages 319–324. Citeseer, 2006.

Møller, M. F. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6(4):525–533, 1993.

Moore, B. C. *Cochlear hearing loss: physiological, psychological and technical issues*. John Wiley & Sons, 2007.

Nawab, S. H., Quatieri, T. F., and Lim, J. S. Signal reconstruction from short-time fourier transform magnitude. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 31(4):986–998, 1983.

Necciari, T., Balázs, P., Holighaus, N., and Søndergaard, P. L. The erblet transform: An auditory-based time-frequency representation with perfect reconstruction. In *ICASSP*, pages 498–502. IEEE, 2013. URL http://dblp.uni-trier.de/db/conf/icassp/icassp2013.html#NecciariBHS13.

Oksanen, S., Parker, J., and Välimäki, V. Physically informed synthesis of jackhammer tool impact sounds. In *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, York, United Kingdom, September 2013.

O'Leary, S. and Robel, A. A montage approach to sound texture synthesis. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 939–943, Sept 2014.

Pearson, K. and Shohat, J. Editorial note to 'inequalities for moments of frequency functions and for various statistical constants'. *Biometrika*, 21(1-4):370–375, 1929.

Portilla, J. and Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int'l Journal of Computer Vision*, 40(1):49–71, December 2000. doi: 10.1023/A:1026553619983.

Průša, Z., Søndergaard, P. L., Holighaus, N., Wiesmeyr, C., and Balazs, P. The large time-frequency analysis toolbox 2.0. In Aramaki, M., Derrien, O., Kronland-Martinet, R., and Ystad, S., editors, *Sound, Music, and Motion*, Lecture Notes in Computer Science, pages 419–442. Springer International Publishing, 2014. ISBN 978-3-319-12975-4. doi: {10.1007/978-3-319-12976-1_25}. URL http://dx.doi.org/10.1007/978-3-319-12976-1_25.

Röbel, A. A new approach to transient processing in the phase vocoder. In *6th International Conference on Digital Audio Effects (DAFx)*, pages 344–349, London, United Kingdom, Septembre 2003. URL http://articles.ircam.fr/textes/Roebel03a.

Röbel, A. Shape-invariant speech transformation with the phase vocoder. In *InterSpeech*, pages 2146–2149, Makuhari, Japan, Septembre 2010. URL http://articles.ircam.fr/textes/Roebel10c/.

Röbel, A. and Rodet, X. Real time signal transposition with envelope preservation in the phase vocoder. In *International Computer Music Conference*, pages 672–675, Barcelona, Spain, Septembre 2005. URL http://articles.ircam.fr/textes/Roebel05a/.

Ruggero, M. A. Responses to sound of the basilar membrane of the mammalian cochlea. *Current opinion in neurobiology*, 2(1873-6882 (Electronic)):449–456, 1992.

Saint-Arnaud, N. and Popat, K. Computational auditory scene analysis. In Rosenthal, D. F. and Okuno, H. G., editors, *Computational auditory scene analysis*, chapter Analysis and synthesis of sound textures, pages 293–308. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1998. ISBN 0-8058-2283-6. URL http://dl.acm.org/citation.cfm?id=285582.285601.

Schwarz, D. *Data-Driven Concatenative Sound Synthesis*. PhD thesis, Ircam - Centre Pompidou, Paris, France, January 2004. URL http://www.ircam.fr/anasyn/schwarz/thesis.

Schwarz, D. State of the art in sound texture synthesis. In *Proc. Digital Audio Effects (DAFx)*, pages 221–231, 2011.

Schwarz, D. and Schnell, N. Descriptor-based sound texture sampling. In *7th Sound and Music Computing Conference (SMC)*, July 2010.

Serra, M.-H. Musical signal processing, chapter introducing the phase vocoder. In *Studies on New Music Research. Swets & Zeitlinger*, pages 31–91, 1997.

Strickland, E. A. and Viemeister, N. F. Cues for discrimination of envelopes. *The Journal of the Acoustical Society of America*, 99(6):3638–3646, 1996.

Strobl, G., Eckel, G., Rocchesso, D., and le Grazie, S. Sound texture modeling: A survey. In *Proceedings of the 2006 Sound and Music Computing (SMC) International Conference*, pages 61–5. Citeseer, 2006.

Tyler, C. W. Beyond fourth-order texture discrimination: generation of extreme-order and statistically-balanced textures. *Vision Research*, 44(18):2187 – 2199, 2004. ISSN 0042-6989. doi: http://dx.doi.org/10.1016/j.visres.2004.03.032. URL http://www.sciencedirect.com/science/article/pii/S0042698904001750.

Verron, C., Pallone, G., Aramaki, M., and Kronland-Martinet, R. Controlling a spatialized environmental sound synthesizer. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, pages 321–324, Oct 2009. doi: 10.1109/ASPAA.2009.5346504.

Welch, B. L. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, pages 28–35, 1947.

Yellott Jr, J. I. et al. Implications of triple correlation uniqueness for texture statistics and the julesz conjecture. *JOSA A*, 10(5):777–793, 1993.

Zhu, X. and Wyse, L. Sound texture modeling and time-frequency lpc. In *Proceedings of the 7th international conference on digital audio effects DAFX*, volume 4, 2004.

Zhu, X., Beauregard, G. T., and Wyse, L. Real-time iterative spectrum inversion with look-ahead. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 229–232. IEEE, 2006.

Zielinski, S., Hardisty, P., Hummersone, C., and Rumsey, F. Potential biases in mushra listening tests. In *Audio Engineering Society Convention 123*, Oct 2007. URL http://www.aes.org/e-lib/browse.cfm?elib=14237.