Multi-scale Scattering Transform in Music Similarity Measuring

Ruobai Wang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Electrical Engineering

University of Washington

2017

Reading Committee:

Les Atlas, Chair

Sreeram Kannan

Program Authorized to Offer Degree:

Electrical Engineering

University of Washington

Abstract

Multi-scale Scattering Transform in Music Similarity Measuring

Ruobai Wang

Chair of the Supervisory Committee:
Professor Les Atlas
Electrical Engineering

Scattering transform is a Mel-frequency spectrum based, time-deformation stable method, which can be used in evaluating music similarity. Compared with Dynamic time warping, it has better performance in detecting similar audio signals under local time-frequency deformation. Multi-scale scattering means to combine scattering transforms of different window lengths. This paper argues that, multi-scale scattering transform is a good alternative of dynamic time warping in music similarity measuring. We tested the performance of multi-scale scattering transform against other popular methods, with data designed to represent different conditions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Prof. Les Atlas to be my advisor, find fundings for me, and give me instructions on research.

I would also like to thank my classmates (David, Scott, Tommy, Eldridge, Brad and Morten) who helped me a lot in research.

I would also like to thank my parents who gave me both economic (for the first year) and mental supports for my study in UW.

Chapter 1. INTRODUCTION

Music similarity measuring is a very important topic. It can be used in music match and recommendation.

Generally, we separate songs into frames, and calculate parameters for each frame. The time-parameter matrix represents the song. The most simple case is to use STFT to calculate the frequency spectrum of each frame, and in this case the matrix is spectrogram. There are many variations of such matrix, while in this paper we will use the term "spectrogram" to represent all of them. Examples of some more advanced methods are MFS (Mel-frequency spectrogram), Chromagram [1], and Scattering transform [2][3].

After we get the spectrogram, we can give a quantitative description on how similar two songs are, by comparing their spectrograms frame by frame. We can use cosine similarity as the similarity measurement, or kinds of distances (such as Manhattan distance [4] or Euclidean distance) to indicate the "dissimilarity".

If two pieces of music are performed according to the same score (sheet music), they should be identified as "very similar". However, different starting time of recording causes a time shift; changing key and tempo also causes translation or zooming in spectrograms; random mistakes or adjustments by the performer also changes how the spectrogram look like.

Dynamic Time Warping (DTW) [4][5] is a method to solve the problem, by trying to align similar frames together without breaking the temporal order. DTW is good at dealing with macroscopic time shift, but not good at dealing with variations in feature domain or within a frame.

Scattering transform, however, is good at dealing with variations within a frame, in both time and frequency domain. It is proved in [2] that, scattering transform is stable under time deformation

of certain level. However, scattering transform cannot deal with non-local time deformations, or to say, very long time shift. Combining scattering with DTW is a solution, which uses the advantage of both methods.

However, in this paper we argue that, multi-scale scattering transform can help solve the non-local time deformation problem too, and therefore we do not need DTW. Multi-scale means that, we do scattering transform with a few different window lengths. Shorter windows provide more information on details, and prevent from marking dissimilar music as similar. Longer windows have much longer "local" ranges, which is a complement to deal with macroscopic time shifts.

In Chapter 2, we will introduce scattering transform and dynamic time warping. In Chapter 3, we will talk about why and how do we use multi-scale scattering transform. In Chapter 4, we will show our experiments on comparing the performance of multi-scale scattering and other methods. In Chapter 5, we will conclude our work, and discuss about what we can improve in the future.

Chapter 2. PREVIOUS WORK

2.1 TIME DEFORMATION AND MEL-FREQUENCY SPECTROGRAM

In chapter 1, we talked about many common phenomena that may disturb the spectrogram, such as time shift, change of tempo, or random note length variation. They are called "time deformation", as they can be described by such a formula:

$$x_\tau(t) = x\big(t - \tau(t)\big) \tag{2.1}$$

Where $x(t)$ is the original signal, $\tau(t)$ is the time-warp function, and $x_\tau(t)$ is the deformated signal.

Mel-frequency spectrum is said to be stable under limited "deformation size" $\sup_t|\tau'(t)|$, as proved in [2]. Suppose we have a signal $x_\tau(t) = x\big((1-\varepsilon)t\big)$, then the deformation size is ε. Each frequency component of $x_\tau(t)$ is zoomed by factor $1/(1-\varepsilon)$. Assume that a harmony was in the k-th frequency bin before time deformation. After deformation, it moves to the $k/(1-\varepsilon)$-th bin, which moved $k\varepsilon/(1-\varepsilon)$ bins. Since $\varepsilon/(1-\varepsilon)$ is fixed and k is increasing as frequency increases, high-frequency components can move to other frequency bins, and since we are doing bin-wise comparing, this frame can no longer be detected to be "similar" as before deformation.

Mel-frequency, however, does not suffer from this. It is a constant-Q spectrum, which means that there are Q frequency bins in each octave. The ratio between frequencies of two consecutive bins is $2^{1/Q}$, or to say, the log-frequency difference is $1/Q$. For time deformation $x_\tau(t) = x\big((1-\varepsilon)t\big)$, the log-frequency is changed by $-\log_2(1-\varepsilon)$, which is $-Q\log_2(1-\varepsilon)$ bins, a constant that does not increase as frequency increases. If ε is small enough, then the frequency components of deformated MFS can still fall into the same bins as the original MFS.

2.2 SCATTERING TRANSFORM

Scattering transform [2] is an extension of MFS. When calculating MFS, we first calculate STFT of the frame. Then, for each frequency bin of MFS, we take those STFT frequency bins that are close to the central frequency of the MFS bin, and do a weighted sum.

We can see that, each bin of MFS has the same effect as a band-pass filter. MFS collects the energy responses of the signal passing through a series of band-pass filters. We can write the filters into time domain, and we get a constant-Q filter bank, which contains one low-pass filter ϕ, and many band-pass filters ψ, where ψ_k is the band-pass filter with the k-th lowest pass band, as shown in Figure 2.1.

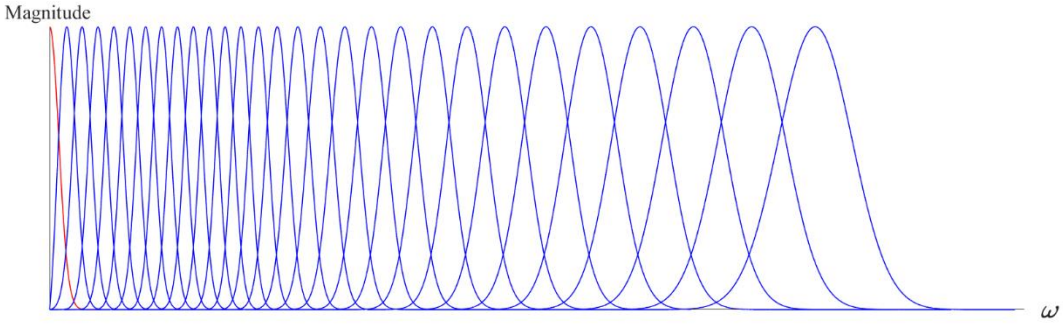

Figure 2.1. Band-pass filters of Mel-frequency spectrum and scattering.

However, as stated in [2], MFS (as well as other spectral methods) needs convolution with a low-pass window function ϕ again after convolution with each filter in the bank, which makes the spectrogram smooth, and loses some high-frequency information.

Scattering transform solves this problem by doing multiple times of filtering. The first layer is just MFS, which gives us $x * \phi$ and $|x * \psi_\lambda| * \phi$. Here $|\cdot|$ means to keep the amplitude and discard the phase of a complex signal. The 2nd layer does two rounds of filtering before the final convolution with ϕ, which gives $\left||x * \psi_{\lambda_1}| * \psi_{\lambda_2}\right| * \phi$. We can do it for even more layers, and get:

$$S_m x(t, \lambda_1, \ldots, \lambda_m) = \left| \left| |x * \psi_{\lambda_1}| * \ldots \right| * \psi_{\lambda_m} \right| * \phi(t) \tag{2.2}$$

The second layer preserves information discarded by MFS, and each deeper layer preserves information discarded by the previous layer.

2.3 Dynamic Time Warping

Dynamic time warping (DTW) [4] is a method to align two certain sequences. DTW works very well if the signal is only warped in time domain, but it is not useful when the signal changes a lot in the frequency domain. DTW works as follow:

For two signals x and y, suppose they have M and N frames. Denote X_m as the m-th frame of signal x, and Y_n is similarly defined. Calculate their spectrograms, and then compare the distance of features of each frame pair: $d(m,n)$ is the distance between frame X_m and frame Y_n. Then we start from the bottom-left point $(1,1)$ of the distance table d, and find an optimal path towards the top-right point (M,N), which minimizes the total distance on the path.

There are only limited kinds of allowed steps. For example, there are only 3 kinds of steps in classic DTW:

1) Go right for 1 block, which makes $(m,n) \to (m+1,n)$, and registers distance $d(m+1,n)$ once. It aligns both frame X_m and X_{m+1} to frame Y_n, and its physical meaning is that y is more shrinked than x at this time point.

2) Go up for 1 block, which makes $(m,n) \to (m,n+1)$, and registers distance $d(m,n+1)$ once. It aligns both frame Y_n and Y_{n+1} to frame X_m, and its physical meaning is that x is more shrinked than y at this time point.

3) Go right and up, each for 1 block, which makes $(m, n) \to (m+1, n+1)$, and registers distance $d(m, n+1)$ twice (because it is two blocks in total). It aligns both frame Y_n to X_m, and frame Y_{n+1} to X_{m+1}. Its physical meaning is that x and y have similar pace at this time point.

Some other types of DTW may allow larger step size, but you can never move left or down, since you cannot break the time order, or to say, the monotonicity condition.

Also, the access to certain areas are sometimes limited, which is called the global constraint. Classic DTW does not have global constraints, but some modified DTW algorithms do. A common example is Sakoe-Chiba band [4], when $M = N$, you can force the path to be around the diagonal, which is described by $|m - n| \leq \varepsilon N$. Another example is Itakura parallelogram, which makes $\min\left\{\frac{m}{n}, \frac{n}{m}, \frac{M-m}{N-n}, \frac{N-n}{M-m}\right\} \geq 1 - \varepsilon$, and it is a thin parallelogram lying on the diagonal. Global constraints are made to eliminate those "false" time warpings, which have low accumulated distance, but warp the signal much too heavier than real-world ones.

It can be seen that, DTW is easily solved by dynamic programming.

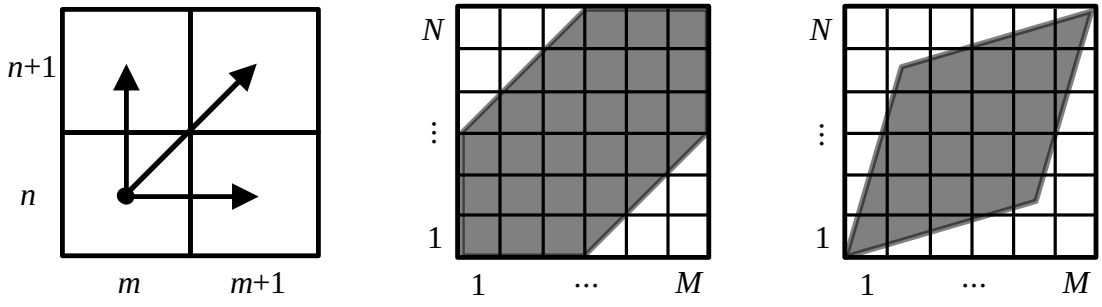

Figure 2.2. Some features of Dynamic time warping.
(Left: Valid steps in classic DTW. Middle: Sakoe-Chiba band. Right: Itakura parallelogram.)

Chapter 3. MULTI-SCALE SCATTERING TRANSFORM

3.1 THE REASON OF MULTI-SCALE

Scattering transform is not good at dealing with non-local time deformations, because it does not align frames like DTW. Given two signals of same length, after scattering transform on each frame, we only compare the frames with the same index, X_n and Y_n.

Suppose that $y(t)$ is the time deformated version of $x(t)$. If the time deformation caused a very long time shift, then the signal in a frame, e.g. X_n, is moved to another frame such as Y_{n+k}, and Y_n is filled with something else. Since we do not compare X_n and Y_{n+k}, we cannot show their similarity; if the time shift is so long that Y_n and X_n are in different notes, we cannot expect to get a high similarity value, though the two signals should be similar.

One solution is to align the frames with DTW, but there is another solution. If we make the time deformation local (comparable with frame length, or even shorter), then there is no problem at all. Though the time deformation is fixed and cannot be shortened, we can increase the frame length!

However, too long frame length causes new problems. First, some notes are short, so one frame can contain many notes. If the notes in each frame are permutated (see section 4.3), we can say that the new music is not similar to the original one; however, the long frame cannot detect it, and gives high similarity value.

Second, scattering transform with shorter frames tend to concentrate the energy in the first and second layer. If the frame goes very long, like ~1000ms, then a lot of energy is spreaded into higher layers [2], as shown in Table 3.. We have to increase the depth (number of layers) to avoid missing this information. However, more layers means much slower running speed and much

higher computer memory requirement, so the compensation between accuracy and complexity is a hard problem.

Table 3.1. Energy concentration in each layer for different window lengths

Energy	Layer 1	Layer 2	Layer 3	Layers 4+
23ms window	94.5%	4.8%	0.2%	0.5%
93ms window	68.0%	29.0%	1.9%	1.1%
370ms window	34.8%	52.0%	11.1%	2.1%
1.5s window	21.0%	47.4%	25.5%	6.1%

Therefore, we need both short frames and long frames. We select a set of scales (window lengths) we need, and do scattering transform of the signal for multiple times, while for each scattering we choose one scale. After that, we normalize the energy, and use all the transform results for similarity evaluation.

3.2 STEPS OF MULTI-SCALE SCATTERING

Let's start with a simple case, where there are only 2 scales.

For 44100Hz sample rate, we can set the short frame to be 93ms (4096 sample points), and the long frame to be 370ms (16384 sample points). It is shown in Table 3. that, for 93ms window, the energy of scattering transform result is concentrated in the first 2 layers, while for 370ms window, it is concentrated in first 3 layers.

Then, we can calculate a 2-layer scattering transform of the signal using 93ms window, and a 3-layer scattering transform of the signal with 370ms window.

Here we have two choices about frame shifts. One is constant-value frame shift, such as 50% of the shortest frame length. In this case, there is about the same number of frames for each scale, no matter how long the frame is; therefore each long frame corresponds to one shorter frame of each scale. However, this option is not preferred, because it makes the frame overlap too large for

long frames (in this simple case, 87.5% for 370ms window), and it also extremely increases the time we spend to calculate the long-window transform.

Then, we choose constant-ratio frame shift, such as 50% of frame length in each scale. Each long frame covers a segment of signal that is represented by many consecutive short frames, in this simple case, each 370ms frame covers 4 of the 93ms frames.

Then, for each shortest frame, we also concatenate its scattering result with those scattering results of the longer frame of each scale that covers this short frame. After that, we use the concatenated frame results for comparing and similarity evaluation.

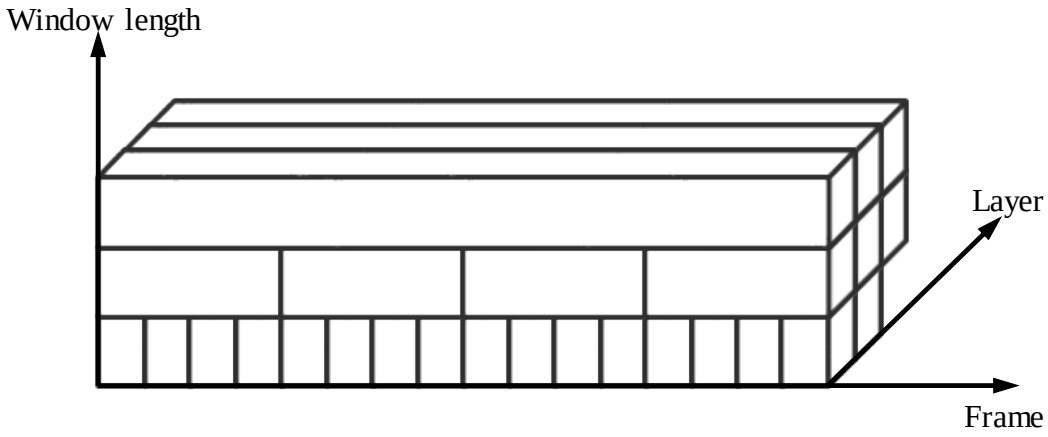

Figure 3.1. The structure of multi-scale scattering.

Chapter 4. EXPERIMENT RESULTS

In this section, we will introduce the experiments details, such as test data, parameter choice, test results, and conclusion for each test. It is worth mentioning that, the Matlab toolbox provided by the authors of scattering transform [2] fixed the max layer to 2, and I modified it to enable the scattering transform of higher layers.

4.1 PERFORMANCE TEST OF SCATTERING ON LOCAL DEFORMATION

Recall formula (2.1) that $x_\tau(t) = x\big(t - \tau(t)\big)$. Denote the scattering transform of x as Sx. It is proved in [2] that,

$$\frac{\|Sx_\tau - Sx\|}{\|x\|} \leq (2 \max Q) \sup|\tau'(t)| \tag{4.1}$$

Suppose we say scattering transform is stable when $\|Sx_\tau - Sx\| \leq \|x\|$, then to make it stable, we need $\sup|\tau'(t)| \leq (2 \max Q)^{-1}$. Since we choose $Q = 8$ (same as MFS) in the first layer, and $Q = 1$ in the higher layers, so $(2 \max Q)^{-1} = 1/16$.

However, it is not necessarily the exact limit, it may still be stable even if $\sup|\tau'(t)|$ is larger. In this section we will investigate how $\sup|\tau'(t)|$ and $\sup|\tau(t)|$ affect the stability of scattering transform.

We are using a single-scale window of 93ms length, and 2 layers of scattering. We choose 2 clips from different songs, and generate 10 time-warped signals from each clip. We start from $\tau(0) = 0$ and a constant $|\tau'(t)|$ value. Whenever $|\tau(t)|$ reaches the limit, the $\tau'(t)$ flips to its opposite number. The original $\tau'(t)$ is uniformly distributed between 0 and the desired $\sup|\tau'(t)|$.

The similarity value is the cosine similarity over the whole scattering spectrogram. Fisher score [6] is Fisher's linear discriminant of this univariate case, the higher score means the better

ability to identify similar and dissimilar pairs correctly. We regard "signals warped from the same clip" as similar, and "signals warped from different clips" as dissimilar.

The result is shown in Table 4.1:

Table 4.1. Similarity values under different warpings

$\sup\|\tau'(t)\|$ among all pairs $/(2\max Q)^{-1}$	$\sup\|\tau(t)\|$ /window length	Similarity value of similar inputs		Similarity value of dissimilar inputs		Fisher score
		mean	stdev	mean	stdev	
100%	50%	0.8373	0.0613	0.3329	0.0054	67.2076
140%	50%	0.7978	0.0901	0.333	0.0043	26.5274
200%	50%	0.7426	0.1027	0.334	0.0055	15.7682
400%	50%	0.6609	0.1195	0.3377	0.0073	7.2842
100%	100%	0.777	0.0549	0.3337	0.004	64.8652
100%	200%	0.7205	0.0569	0.3371	0.0049	45.0041
100%	400%	0.6906	0.0711	0.3437	0.0074	23.5793

We can see that, increasing $\sup|\tau'(t)|$ makes the result worse in all aspects, especially a decrease of the mean similarity value, and an increase of the variance, of similar pairs, which results in a quick drop in Fisher score.

Increasing $\sup|\tau(t)|$ only affects the mean of similar pairs, and the mean drops slowly. Therefore, we can say that $(2\max Q)^{-1}$ is a good limit for warping rate. We can also see that, scattering transform is still quite stable under "semi-local" time deformation, which has a maximum time shift of a few frame lengths.

4.2 Performance Comparison on Local Deformation

Here we conduct another test on different kinds of similarity evaluation methods. They are:

1) Multi-scale scattering (2-layer 93ms scatter + 3-layer 370ms scatter)

2) Long window scattering (3-layer 370ms scatter)

3) Scattering (93ms window, 2 layers)

4) MFS (93ms window) with DTW

5) STFT (93ms window) with DTW

The signal is generated the same way as in Section 4.1, and the length of each clip is 10^6 sample points (22.675 sec). The max warping rate is $(2 \max Q)^{-1} = 1/16$, and the max time shift is 185ms. Result is shown in Table 4.2.

Table 4.2. Similarity values of different methods

Method	Similarity value of similar inputs		Similarity value of dissimilar inputs		Fisher score
	mean	stdev	mean	stdev	
Multi-scale scattering	0.7763	0.0561	0.3630	0.0048	53.8293
Long window scattering	0.8839	0.0618	0.4130	0.0054	57.5886
Scattering	0.7205	0.0569	0.3371	0.0049	45.0041
MFS+DTW	0.9230	0.0443	0.6599	0.0120	32.8364
STFT+DTW	0.6786	0.1179	0.3652	0.0166	6.9219

We can see that, scattering has much better performance than DTW. Even combined with MFS, DTW cannot outperform scattering, which convinces us that scattering is a better choice than DTW in local deformation circumstances.

Multi-scale scattering seems to be a combination of short-window and long-window scattering. We can expect it to be more robust than single-scale, if each scale has its limitations in practice. Here we have seen an example of long window better than short window. In next section we will see why we still need shorter windows.

4.3 The Importance of Short Windows

At the end of Chapter 1 we introduced that, shorter windows can provide more information on details, and can prevent from marking dissimilar music as similar. Here we test it with an experiment.

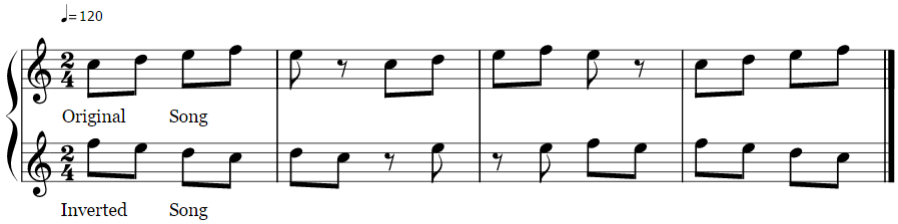

Figure 4.1. An example of how we reverse a song.

In this test, the two clips come from the same song. One clip is directly cut from the original song. Then, to create the other clip, we separate it into sections, each with 4 eighth notes. Then we reverse the order of eighth notes in each section, and concatenate them together in the original order, to get the second clip, as shown in Figure 4.1.

This time we use a song of 22050Hz sample rate, so that each flipping segment has length comparable with the long window. All the other parameters are the same as in Section 4.2. The result is shown in Table 4.3.

Table 4.3. Similarity values in section-reverse test

Method	Similarity value of similar inputs		Similarity value of dissimilar inputs		Fisher score
	mean	stdev	mean	stdev	
Multi-scale scattering	0.9106	0.0188	0.8333	0.0106	12.8230
Long window scattering	0.9662	0.0134	0.9396	0.0119	2.1846
Scattering	0.8882	0.0221	0.7906	0.0100	16.1739
MFS+DTW	0.9448	0.0192	0.8735	0.0080	11.7295
STFT+DTW	0.7890	0.0263	0.7057	0.0149	7.5974

We can see that, this time short-window scattering is much better than long-window ones. Multi-scale scattering, as a combination of short and long-window scattering, keeps its stable performance which is balanced and closer to the better side. Also, multi-scale scattering is still better than DTW-based methods.

Here we have proved by experiment results that, multi-scale scattering is a good alternative of DTW. Also, multi-scale is very important, because it has the advantage of both short and long scales, making it able to deal with rather long time shift, while still being able to detect the short-range dissimilarities.

Chapter 5. DISCUSSION AND CONCLUSION

We can see that, multi-scale scattering is a competent method in music similarity measuring, It has a better performance than dynamic time warping, on those time deformations with moderate time shift but affect the frequency domain very much. Both large and small scales are important, because long window length is needed to deal with macroscopic time shifts that are "non-local" for short frames, while short frames can detect the small dissimilarities that the long frames cannot. Therefore we verified that multi-scale scattering transform is an important method in similarity measuring.

However, there are still some things we can do. First, we need to test its performance on real-world music and songs, where "similar" songs can have similar style or melody, but very different details. Second, the running speed is not fast, and the memory cost is huge, which needs to be improved if possible. Third, when the time shift is so long that it overwhelmingly exceeds the range of the longest window scale we can use, simple scattering transform may not work as desired, so we may need to search for pre-alignment, which is not a problem for DTW.

If these problems are answered, multi-scale scattering can be more promising and useful in solving practical problems.

BIBLIOGRAPHY

[1] Bartsch M A, Wakefield G H. Audio thumbnailing of popular music using chroma-based representations[J]. IEEE Transactions on multimedia, 2005, 7(1): 96-104.

[2] Andén J, Mallat S. Deep scattering spectrum[J]. IEEE Transactions on Signal Processing, 2014, 62(16): 4114-4128.

[3] Rodríguez-Algarra F, Sturm B L, Maruri-Aguilar H. Analysing Scattering-Based Music Content Analysis Systems: Where's the Music?[J]. 2016.

[4] Müller M. Information retrieval for music and motion[M]. Heidelberg: Springer, 2007.

[5] Orio N, Schwarz D. Alignment of monophonic and polyphonic music to a score[C]//International Computer Music Conference (ICMC). 2001: 1-1.

[6] Mika S, Ratsch G, Weston J, et al. Fisher discriminant analysis with kernels[C]//Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop. IEEE, 1999: 41-48.

VITA

Ruobai Wang is a 2nd year graduate student in department of Electrical Engineering, University of Washington. He got his Bachelor's degree in Tsinghua University, China. He is also planning to get a Master's degree during his PhD study.

Ruobai started to do research about audio signal processing in 2014. After he came to UW, he has also explored many other research fields. At the end of 2016, he decided to focus on audio signal processing again, with other signal processing and machine learning skills he learned these years.