Interpretable Parameters for Timbre Analysis and Synthesis

Han Zhang

Abstract

Timbre is a mysterious property of music that is so important in music but still remains complicated to describe in a quantifiable way. Some studies on timbre stood upon a subjective perspective and yield semantic descriptors, while some others, on the other hand, treated music as a kind of signal without paying much attention to the perceptual characteristics. Considering that both the subjective understanding and the information in data are equally important, this work proposes a middle space of interpretable parameters in between that takes the advantages from both sides. In this documentation, the parameter extraction model is explained in detail and the informativeness and descriptive capability of the space is verified by a musical instrument recognition experiment. On the synthetic side, some methods for timbre design will be specified, and a GUI[1] for interactive demonstrations of the work is also implemented. Finally, some discussions are made with the expectation for a prosperous development of the approach of designing sound and music directly by sculpting the morphology of the spectrogram. [2]

1 Introduction

Timbre is an attribute of musical sound that have always been interesting for researchers, but hasn't been well-defined for decades. Many efforts have been made to find a proper description or a way of quantification, but there is still not a widely agreed numerical definition for timbre until now. One reason is that timbre generally plays the role of a complementary property of pitch, loudness, and duration[1], which has a broad sense of sound characteristics. Therefore, a lot of studies from various standing points were proposed and promoted regarding the describing and synthesizing of timbral sound.

In early studies, many attentions naturally focused on standardizing semantic descriptors[2][3][4] and finding multi-dimensional timbre space[5][6][7].These works all have a common underlying emphasis on the fact of human perception, for which they derive their result from human listening tests. In addition to defining some quasi-orthogonal dimensions, they also tried to find quantifiable scales for the axis. However, these two requirements are hard to be met at the same time. Other later approaches involve the perspectives of signal processing, which treats music timbre as a audio signal and applies models in audio processing to discover the magic behind. Sound have two common forms of representation, the serial representation as a raw waveform, and the matrix form as a spectrogram. Research in this field actively use both as the stimuli to their models. The former, due to its high complexity and large data amount, always needs algorithms that have the power to extract some high-level features, thus is preferred by some deep learning methods[8][9]. The later, involving a transforming of data at first, though is equally informative in machine learning models[9][10], it also provides possibilities for traditional signal processing research[11][12].

This orientation in methodology exploits the state-of-art models and without a doubt has more potentials in digging out and manipulating the information. However, it nevertheless neglect some subjective aspects which leads to some difficulties in the interaction with users who tend to modify or design timbre of their own interest. To both take the advantages of the data and preserve the interpretability of the controlling features, this work defined a parameter space with a set of features extracted from the spectrogram of a sound that is at the same time understandable. It is always possible to reconstruct the sound from the parameters, which verify the fidelity of the information on the parameter space. I also verified the interpretability and the distinguishability by designing

[1] https://github.com/ZhangHanpqqo/timbre_analysis_synthesis

[2] Presentation of the work: https://zhanghanpqqo.github.io/HanZhang/assets/HanZhangThesis.pdf

experiments to map the parameters to existing semantic descriptors and to recognize the identity of musical instruments. To test the potentials of synthesizing sounds, I also tried sound morphing, which is the interpolation of two different timbre, in the parameter space. Finally, I implemented a GUI to allow real-world interactions between the data and human.

This paper is organized as follows. Section 2 introduces some works that are closely related, though not exactly the same, to this topic of analysing and synthesizing timbre from the spectrogram. Section 3 explains details of the parameter extraction model which contains the fundamental models, description of the parameters and the discussion on the reconstruction results. Section 4 introduces the model's capability in timbre analysis with the experiment on instruments recognition. Section 5 elaborates some possibilities of the synthetic power of sound modification and morphing. In section 6, a conclusion of this work is made, and finally in section 7, some outlooks of the future works is stated.

2 Related Works

2.1 Spectral Modeling Synthesis

Many sound synthesis models are based on the fact that sound can be decomposed by a combination of some time-varying sinusoidal waves with various frequency[13], which is commonly named the additive model of audio. This model has tested to be informative and expressive by many vocoders, and the Spectral Modeling Synthesis[11] tool is a relatively comprehensive model for detecting the deterministic sine components and modeling the residual noise component with stochastic model. It shows the ability of extracting harmonics by tracking the frequency of amplitude peaks within each time frame and concatenating the results, and further reconstructing the sound by adding up the sine waves according to the frequency and amplitude detected. It gives good performance not only for pitched musical sounds, but also for more complicated unpitched percussive sound. Among the several models it provides, the Harmonics plus Stochastic model gives a reliable detection of the sinusoidal component and a concise representation of the rest, and it is the embedding model used in the parameter extraction model in this work which is explained in later sections.

2.2 AudioSculpt

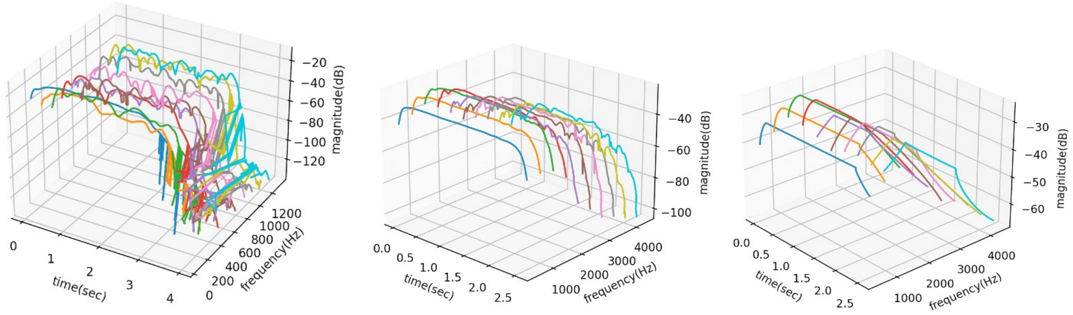
AudioSculpt[14] is a commercial software developed by IRCAM, France. It is a sound processing software and the processes happen under some graphic represents of sound including spectrogram, and it maturely integrates tools for sound producing by directly drawing on the spectrogram. The reason for mentioning it here is to compare this work with some existing products that share similar means of representing the sound and process but may have different ambition. While AudioSculpt relies on a imported sound to be modified as a source, this parameterization approach contains the power of generating a sound from the scratch, and the structured form of data makes the analysis steps easier than unstructured data types like the spectrogram. Even with the difference, the explicitness of graphic representations in AudioSculpt still inspired the parameters setting section of the GUI for this work.

3 Parameter extraction model

For most pitched musical timbre, most of its energy falls in the harmonics and the beginning part where the attack happens, and they are perceptually important as well, since the cochlea receive vibration by frequency, and the level of the stimuli depends on the energy in each critical band[15]. The main focus of this work is to use features that reflect the morphological characteristics of the harmonics. The attack, which is also significant for the identification, is always considered as a separate problem due to its own complexity and is not particularly discussed in this paper.

3.1 Additive model

The additive model is at the heart of the whole system. Generally, a audio signal $x(t)$ can be expressed as the summation of a set of sinusoidals multiplied by time-varying weights for each frequency band,



(a) (b) (c)

Figure 1: 3-dimensional plotting of the first 10 harmonics derived from a trumpet playing in A4 in different steps. The three dimensions are time, frequency, and magnitude. (a) Harmonics extracted by HpS tool. (b) Smoothed harmonics. (c) Harmonics reconstructed from the parameters.

namely:

$$x(t) = \Sigma_{k=1}^{N} A_k(t)sin(\omega_k(t) * t + \phi_k(t)) \tag{1}$$

where $A_k(t)$ means the amplitude for the k-th harmonic at time t, and $\phi_k(t)$ is the phase offset.

For an acoustic signal, the amplitude for the magnitude and the phase can be easily calculated with the help of time-frequency transforming strategy, like the Short Time Fourier Transform(STFT). There are arguments that the phase is not perceivable for musical timbre and can be randomized, but some works deliberately test on the effect of phase and assess the importance of the phase[16]. Considering the potential significance of phase especially in sounds with larger transients, we still need to take the phase spectrogram into account though it's not necessarily to be accurate.

The Harmonics plus Stochastic(HpS) model[11] utilized as a fundamental harmonics detection method in this work is based on the additive model as well. The harmonics detected by the HpS can be visualized in a 3-dimensional time-frequency-magnitude space shown in Figure 1(a).

3.2 Parameters

There are three time-varying attributes appear in the model: Frequency, Magnitude and Phase, thus every harmonic indeed can be controlled given these three properties. Before separately viewing the details of the parameters related to these three dimension, it is worthwhile to mention that all the harmonics are denoised by eliminating the points where harmonics are not successfully detected, i.e. the sharp stitches in Figure 1(a), and assign values by doing interpolations before getting parameters from. This pre-process practically reduces the disturbing effects caused by the false detection thus credibly increases the reliability of the parameters. The denoising result is visualized in Figure 1(b), in which the harmonics are obviously smoother than that of (a).

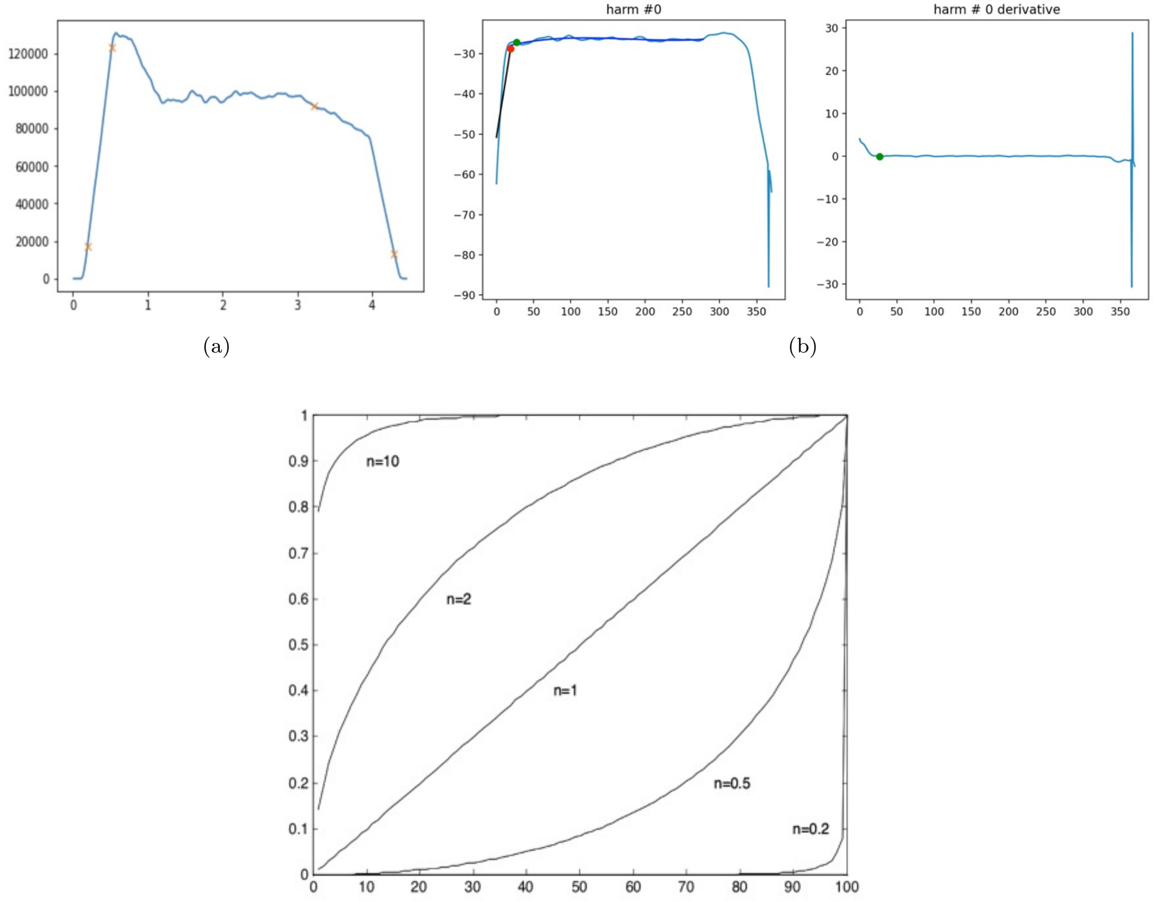
3.2.1 Frequency

It is not hard to observe that the frequencies are not consistent along the time frames. Some random fluctuations in the peak frequencies are slightly altering from frame to frame. More importantly, all kinds of musical instruments are not purely harmonic due to the vibrating nature of its acoustic structure.In other words, the higher partial frequencies are not necessarily to be the multiples of the fundamental. This inharmonicity is called the stiffness of a sound[17]. Although some literature suggest that the stiffness is not audible[18], it is still considered valuable to save the characteristic for analytic reason and for scaling to rougher non-musical timbre later.

To represent the frequencies, we use the mean and variance for each harmonic's frequency offset from the exact multiplication of the fundamental with an underlying observation that the offsets are distributed in a Gaussian manner. When doing reconstructions, the frequencies are first allocated according to the distributions, following which is a smoothing process, for the random fluctuations can bring in extra level of high frequency noise when testing.

3.2.2 Magnitude

Magnitude jointly contains two piece of information: the energy distribution across frequency and the internal magnitude changes over time. The energy distributions rarely change within one note for most of the instrumental sound, and there is a widely accepted Attack-Decay-Sustain-Release(ADSR) model[19] that fits for a large portion of sounds' propagation through time. At the same time, it is equally important to keep the conciseness and the interpretability. Therefore, the scheme is to segment each harmonic with four key points that locate the Start of Attack(SOA), End of Attack(EOA), Start of Release(SOR), and End of Release(EOR). Figure 2(a) shows a typical singular harmonic with a complete ADSR shape being segmented with four key points marked as orange crosses. At the sound modeling stage, it is more efficient to think of the general shape and not look too close the specific fluctuations within the segment, since it only depends on the articulation of the note, but does not have a lot of effects on timbre itself.



(a)

(b)

(c)

Figure 2: (a) Typical ADSR shape of singular harmonic. The crosses mark the four key points of segmentation strategy. From left to right: Start of Attack(SOA), End of Attack(EOA), Start of Release(SOR), and End of Release(EOR). (b) Example on how to ensure the location of EOA. (c) Curve shapes for 2 with various n.

Intuitively, all the points can be detected by some assigned thresholds. For instance, the SOA and the EOA are at the time epoch where the amplitude meets certain ratios of the maximum amplitude of the harmonic if scanning the harmonic from the beginning; likely, the SOR and the EOR can take the same strategy but scan the harmonic reversely. This method works well for SOA, SOR, and EOR, which always give reasonable location on the harmonic with the threshold of 10% of max, 70% of max, and 10% of max. The only exception is the EOA. In a stereotypical impression, the end of attack should be the place with the highest magnitude, so the EOA should be very close to the summit thus has a high threshold for detection. However in real world cases, many recordings in the database contains a natural crescendo, that is a slow rising in volume, within one note, which causes both the maximum point and the EOA turn out to be very close to the end of the note, as the shape shown in the left plot of Figure 2(b). This false detection is unacceptable since the rise time of a sound is a crucial factor for timbre identification[20], thus needs a precise determination.

Therefore, a more delicate method is proposed with a two-step estimation of the precise EOA location. In stead of simply checking the value, it tracks the gradient. The first step is a coarse estimation. After smoothing the harmonic, we calculate the derivative and find the first point whose derivative falls roughly around zero. In the example in Figure 2(b), the derivatives are shown on the right and the green dots in both plots indicate the point found in step 1. Then the harmonic between the SOA and the SOR can be cut into two parts by the point. The second step is a fine tuning step which fits the two parts with a couple of cubic splines and calculate the intersection of the curves to be the EOA. The left plot in Figure 2(b) shows the splines and found their intersection at the red dot, which is designated as the EOA of this harmonic.

Besides the time epoch and magnitude of the key point, the shapes of each segment also need to be recorded. Taking down all the values in the segments can be so redundant and undesirable, but simply doing interpolations between the key points will also lose expressiveness since many shape information will be abandoned. As a result, a single-degree-of-freedom curve which can fit both the concave and convex situation is a suitable compromise. The curve equation being used in this work is:

$$y = y_s + (y_e - y_s)(1 - (1 - \frac{x - x_s}{x_e - x_s})^n)^{\frac{1}{n}} \tag{2}$$

where (x_s, y_s) is the time and magnitude tuple of the key point on the left of the segment and (x_e, y_e) is the tuple of the right key point. Some examples of the curves are plotted in Figure 2(c). The range for the parameter n is from 0 to 40, and the larger n is, the sooner the magnitude approaches the end value.

With the set of parameters representing the location of the key point, the value of the key points, and the shape between them respectively, there is a straightforward path for doing reconstruction. First load the key points for every harmonics, and then calculate the values in between referring to the shape parameter and the identical curve equation. The whole magnitude spectrogram can be covered and being controlled in a relatively small set of data. The reconstruction result of the example trumpet sound is displayed in the 3D plot in Figure 1(c).

3.2.3 Phase

As stated in previous sections, phase is a minor feature in steady musical sound without abrupt transients. It generally follows the propagation principle of sinusoidal waves:

$$\phi_{i+1} = \phi_i + \frac{2\pi \frac{f_i + f_{i+1}}{2} H}{f_s} \tag{3}$$

in which ϕ_i is the phase of the i-th time frame, f_i is the frequency of it, f_s is the sampling frequency, and H is the hop size for the STFT transformation that operates before extracting the harmonics. This equation illustrates that the method to get the phase of the current frame is to add the angle that the sine wave traveled from the last frame to the present.

Now think of the problem backward, we need a take-off line for the propagation, that is the phases for the first frames. It is hardly audible to randomize the values, but a linearity of the phases to the number of harmonics shows in the experiments, so it is reasonable to add two more parameters, the slope and the intercept of the linear regression to anchor the first frame phases and further develop the phase spectrogram.

3.2.4 Residual noise

The residual is the portion of the spectrogram that is not included in the deterministic part of harmonics. This part is mainly composed by some noise and super high partials that are not detected and thus being considered as noise as well. In the HpS model, the level of the residual noise is measured in a much wider frequency band and are sampled much sparser in time. Based on the levels, some stochastic noise can be generated and attached to the corresponding region of sound. Modeling noise can be another complicated problem, so here the original parameters for the stochastic model in HpS are preserved.

3.3 Discussions

Now in the parameter space, the frequency, the magnitude, and the phase all have their own descriptors and can jointly reconstruct a sound. Have to admit that the vividness is sacrificed and some artifacts are also introduced by the parameterization scheme, but many of the degraded consequences can be make up by some post modulation steps. However, there is an advantage that should not be overlooked, which is the conciseness of the space and the interpretability for every single feature.

In Table 1 below all the parameters for the harmonics are listed and the total amount of the parameters is also calculated. The size of the parameter space depends on the harmonics required or can be included under the sampling rate. Usually, detecting 40 harmonics is far beyond enough, and it only consumes 602 parameters in this model, and the size is independent to the duration of the original sound. Comparatively, if using full data for the harmonics of a 1 second sound as those projects using spectrograms as stimuli, also demanding 40 harmonics, then there will be 3480 features to deal with. If the length of the sound keep expand, then the number will grow even larger.

Parameter Name	Parameters amount (N = number of harmonics)
Frequency offset mean	N
Frequency offset variance	N
Magnitude key point time	4N
Magnitude key point magnitude	4N
Magnitude segment shape	5N
Phase of first frames	2
Total	**15N+2**

Table 1: All of the parameters for harmonics and their numbers.

The interpretability is another advantage. Although its benefits are not quantifiable, it is not hard to understand that with the features that can be elaborated with natural language, and that all the changes they bring to the timbre are meeting the expectations, users can easily make use of the model for more explorations in sound designing and making music.

4 Instruments recognition

The feature extraction model presented in the previous section have introduced the composition of the parameter space. In this section, we measure the ability of the space to describe and represent timbre by verifying its capability in musical instruments classification.

Musical instruments recognition is a problem that has arisen many discussions. Settings of this problem have being changing in different aspects, from family-level classification to instrument-level recognition[21], from monophonic input to polyphonic inputs[22][23], from clip-wise resolution to frame-wise resolution[24]. For the classification task in this work, we choose a fairly simple scenario, doing clip-wiser instrument-level multi-class classification on monophonic recordings, and we used a baseline level model to tackle with the problem. It is because the purpose of this experiment is not for designing models to well-perform the mission, but to demonstrate the parameters' capability in analysis timbres. If faced with more completed settings later, it would be feasible to increase the complexity of the classification model.

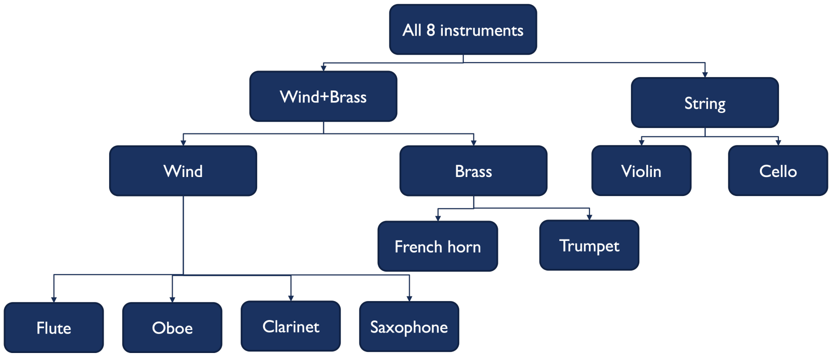


Figure 3: Hierarchical strategy for 8-class musical instrument recognition.

4.1 Model

The task is to classify eight different instruments selected from three orchestra classes. All the instruments and families will be specified in later sections. The model being used here is a random forest with 80 decision trees. Random forest is often used as the baseline model for machine learning tasks, so if the result is on an acceptable level for a multi-class classification, then it is reasonable to believe that with a better tuned model, the parameter space is very likely to provide better performance.

Comparing to directly apply 8-class classification to the data set, it sometimes works better to do it hierarchically, for which we do a family-level classification first and within each family identify each instrument with a subsequent classification. The structure of this hierarchical scheme is shown in Figure 3.

4.2 Database

The recordings we used to train and test the model are from the sound sample library of London Phiharmonia Orchestra[25]. We collected samples for all the eight instruments: flute, oboe, clarinet, french horn, trumpet, saxophone, violin, and cello from their corresponding instrument families: woodwind, brass, and string. There are 1834 samples in train, validation, and test set in total with an approximately equal distribution for each instrument. All the samples are clearly labeled by the name of the instrument, the pitch, the dynamics, the articulation, and the duration. In this task, only steady state articulation are included and recordings with impulsive articulations are left out, like the pizzicato for strings.

The data is randomly split into the training, validation, and test set with a proportion of 70%, 20%, and 10%. The distribution of the instrument, dynamics and duration are approximately identical.

4.3 Result

The accuracy of the recognition for each instrument and the overall accuracy is listed in Table 2. The result, from a general perspective, is sufficient to demonstrate the descriptive power in the parameters. For the two classification strategies, the hierarchical structure gives a better overall result than direct classification, but its effects to individual instruments varies. Most samples detected as woodwinds are correct, and the two instruments in brass family but have close relationships with woodwinds, the french horn and the saxophone also have plausible accuracy. However, the accuracy for trumpet and violin is a little bit disappointing, which indicates that other instruments are very easily to be falsely detected as these two. Need to mention that sometimes it is trick to distinguish the timbre of instruments in same family, especially for string instruments whose timbres sound similar even to well trained musicians. The classification result shows the same fact that for the binary classification between violin and cello, if the pitch is not specified, then the overall accuracy is barely 0.68.

4.4 Analysis

The instruments recognition experiment not only verifies the analytic power of the model, but also provides an angle to rethink of the parameter space itself. In the model of random forest, there is

Instrument Name	8-class	Hierarchical
flute	0.92	0.96
oboe	0.78	0.80
clarinet	0.95	0.92
french horn	0.89	0.84
trumpet	0.60	0.69
saxophone	0.89	0.96
violin	0.53	0.45
cello	0.96	0.83
Overall	**0.79**	**0.81**

Table 2: Classification accuracy for 8 classes of instruments.

a measurement called Gini index[26] that reflects the significance of the features. Intrinsically, this scalar calculates how much impurity one feature is able to reduce, so the importance of a feature in representing the timbre is proportional to its Gini index value.

The importance of parameters are illustrated in Figure 4 and the descriptions for each region are specified at the top. The spikes of data appear periodic manner with the interval of 40, which is the total harmonics detected. In other words, the lower harmonics tend to contain more descriptive information than higher partials.

Among all the parameters, the features that imply the proportions of energy distributed in each harmonic show the greatest importance, and this is not a unique phenomenon for low harmonics since the overall level of significance is also higher than other features. This fact corresponds with the insights of research on timbre descriptor, many of which mention that the descriptors for energy distribution across frequency bands usually are likely to be more obvious and audible to listeners[20]. Other features that are relatively more useful are the rise time, release time, and some features designating the shape of the harmonics.

5 Capability in sound synthesis

In additional to the descriptive function of timbre, another primary purpose of the project is to propose a new approach of sound synthesis by shaping the harmonics with its morphological features. By giving proper values to the parameters, it doesn't take much effort to design a sound from the scratch. The only concern is that it needs some tuning of the parameters to approach the desired sound due to the lack of experience and efficient techniques of tuning. Also, in terms of testing the capability in synthesis, to avoid the subjective factors in evaluating the quality of the sound, generating timbre from the scratch may not be an ideal option.

Instead, importing some reference sounds and use their parameters as templates for tuning would be a better way of testing the possibilities in doing sound synthesis. In the GUI that we implemented for demonstration(sources attached to the footnote of page one) in Figure 5, after importing the reference sound, all the parameters will be extracted automatically and listed in the parameter tuning section on the right. By assigning new numbers or dragging the key points, the parameter and the shape that displays in the widgets on the left will alter accordingly, and the users can always listen to the synthetic sound.

Another interesting attempt is to modify timbre with the characteristics of one another, which we call it sound morphing. Technically, sound morphing means to do interpolations between two given timbres. It can be complicated when using raw waveform or spectrogram as stimuli, so the methods dealing with this problem often involve deep learning models[3]. This model, nevertheless, is naturally convenient for interpolations, since all the features are scalars and have semantic meanings. Tests of doing interpolations in the parameter space give perceptually reasonable result. For example, we

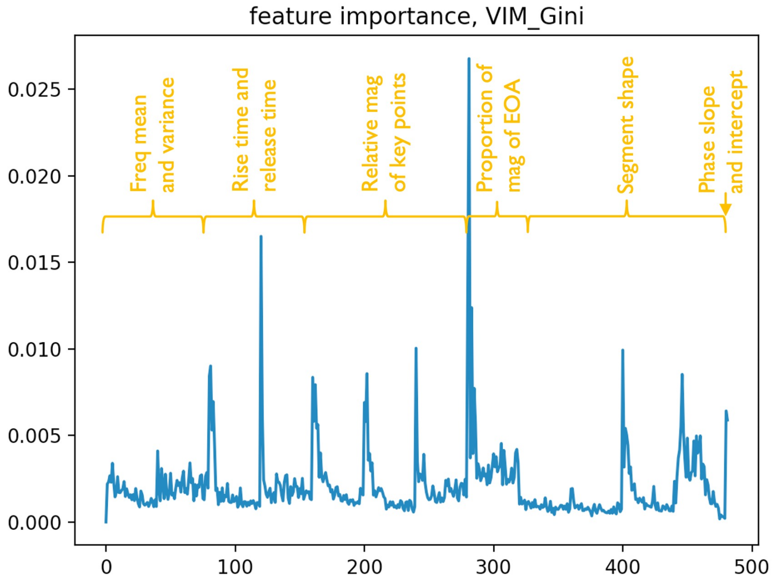


Figure 4: Parameter importance measured by Gini index in the random forest model. The horizontal axis is the number of parameters, and the descriptions for each region are elaborated above the data.

tried to morph a bright trumpet sound with a hollow clarinet timbre given a morphing rate of 0.5, which means we are expecting a sound in the middle of these two timbres, and the system generated a french horn alike timbre which has some brassy factors but is dimmer and less directional. If try to classify the sound with the trained classifier introduced in previous sections with a morphing rate increasing from 0 to 1, then the recognition result will shift from trumpet, oboe, french horn, and finally to clarinet. The trajectory matches the impressions to the timbre evolution. This function is also integrated in the GUI, and a video demo can be found in the URL at first page's footnote as well.

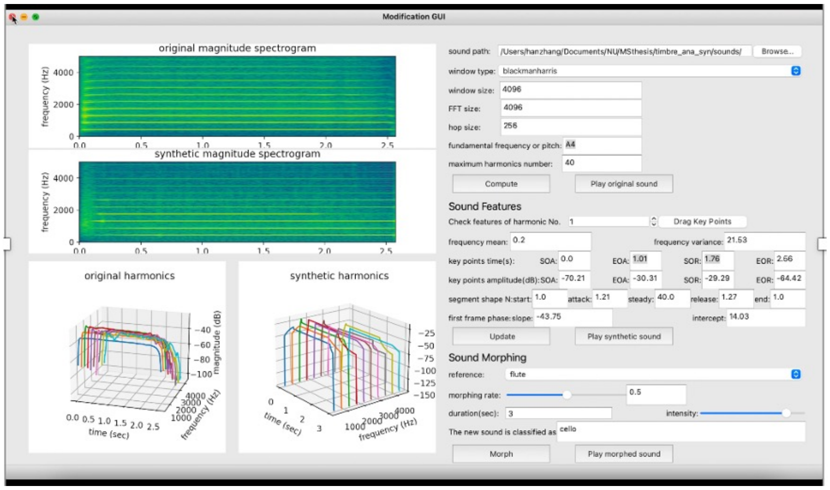
6 Conclusion

The main contributions of the work is to define the understandable parameter space for the morphological characteristics of harmonics in musical sounds, which plays the role of an interpretive bridge between the perceptual nature of timbre and the analytical power of signal processing methods. Its informativeness and its descriptive capability is verified by a musical instrument recognition experiment. On the synthetic side, the parameters also show a good control for timbre modifications. The GUI of the model provides an interactive platform for users to test the functions and play with the parameters for more understanding.

7 Future works

Currently, I have just finished a basic system for morphological parameter extraction and timbre analysis and synthesis based on the parameters. This work just include some fundamental versions for all of the functions being expected and is only sufficient to demonstrate the rough idea of the new approach for designing timbre and music from the spectrogram. This is not a new concept in the music world, but the technological implementation is still behind the creative mind of musicians, so accomplish a complete system for this idea is the destination of this project.

Obviously, it is a huge narrative, so some near future plan is focusing on these following aspects: 1. Refining the parameter extraction model for more precise descriptions. 2. Developing a better attack and residual noise modeling method so that the reconstruction quality of impulsive sounds can be improved. 3. Thinking of modulations and transformations of the reconstructed sound that makes the sound more vivid and explores more possibilities. 4. Mapping the parameters to the semantic descriptions for a better explanation of the acoustical function of the parameters.



sound path: /Users/hanzhang/Documents/NU/MSthesis/timbre_ana_syn/sounds/

window type: blackmanharris

window size: 4096

FFT size: 4096

hop size: 256

fundamental frequency or pitch: A4

maximum harmonics number: 40

Sound Features

Check features of harmonic No. 1

frequency mean: 0.2 frequency variance: 21.53

key points time(s): SOA: 0.0 EOA: 1.01 SOR: 1.76 EOR: 2.66

key points amplitude(dB): SOA: -70.21 EOA: -30.31 SOR: -29.29 EOR: -64.42

segment shape N: start: 1.0 attack: 1.21 steady: 40.0 release: 1.27 end: 1.0

first frame phase: slope: -43.75 intercept: 14.03

Sound Morphing

reference: flute

morphing rate: 0.5

duration(sec): 3 intensity:

The new sound is classified as cello

Figure 5: A screen shot of the GUI for the demonstration of the system. On the left are some widgets displaying the spectrograms and harmonics for both the original sound and the synthetic sound. On the right are the parameter displaying and tuning sections that integrates the sound synthesis functions that are mentioned in Section 5.

References

[1] Jean-Claude Risset and David L. Wessel. “Exploration of Timbre by Analysis and Synthesis”. In: *The Psychology of Music* (1999), pp. 113–169. DOI: 10.1016/b978-012213564-4/50006-8.

[2] Geoffroy Peeters et al. “The Timbre Toolbox: Extracting audio descriptors from musical signals”. In: *The Journal of the Acoustical Society of America* 130.5 (2011), pp. 2902–2916. ISSN: 0001-4966. DOI: 10.1121/1.3642604.

[3] Marcelo Caetano and Xavier Rodet. “Automatic timbral morphing of musical instrument sounds by high-level descriptors”. In: *International Computer Music Conference, ICMC 2010* (2010), pp. 254–261.

[4] Wei Jiang et al. “Analysis and modeling of timbre perception features in musical sounds”. In: *Applied Sciences (Switzerland)* 10.3 (2020). ISSN: 20763417. DOI: 10.3390/app10030789.

[5] David L. Wessel. “Timbre Space as a Musical Control Structure”. In: *Computer Music Journal* 3.2 (1979), p. 45. ISSN: 01489267. DOI: 10.2307/3680283.

[6] Anne Caclin et al. “Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones”. In: *The Journal of the Acoustical Society of America* 118.1 (2005), pp. 471–482. ISSN: 0001-4966. DOI: 10.1121/1.1929229.

[7] John M. Grey. “Multidimensional perceptual scaling of musical timbres”. In: *Journal of the Acoustical Society of America* 61.5 (1977), pp. 1270–1277. ISSN: NA. DOI: 10.1121/1.381428.

[8] Aaron van den Oord et al. “WaveNet: A Generative Model for Raw Audio”. In: (2016), pp. 1–15. URL: http://arxiv.org/abs/1609.03499.

[9] Chris Donahue, Julian McAuley, and Miller Puckette. “Adversarial audio synthesis”. In: *7th International Conference on Learning Representations, ICLR 2019* (2019), pp. 1–16.

[10] Jesse Engel et al. “DDSP: Differentiable digital signal processing”. In: *arXiv preprint arXiv:2001.04643* (2020).

[11] Xavier Serra and Julius O. Smith. “Spectral modeling synthesis. A sound analysis/synthesis system based on a deterministic plus stochastic decomposition”. In: *Computer Music Journal* 14.4 (1990), pp. 12–24. ISSN: 01489267. DOI: 10.2307/3680788.

[12] Philippe Guillemain and Richard Kronland-Martinet. “Characterization of acoustic signals through continuous linear time-frequency representations”. In: *Proceedings of the IEEE* 84.4 (1996), pp. 561–585. ISSN: 00189219. DOI: 10.1109/5.488700.

[13] R. McAulay and T. Quatieri. "Speech analysis/Synthesis based on a sinusoidal representation". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34.4 (1986), pp. 744–754. DOI: 10.1109/TASSP.1986.1164910.

[14] Niels Bogaards, Axel Roebel, and Xavier Rodet. "Sound Analysis and Processing with AudioSculpt 2". In: *International Computer Music Conference (ICMC)*. cote interne IRCAM: Bogaards04a. Miami, United States, Nov. 2004, pp. 1–1. URL: https://hal.archives-ouvertes.fr/hal-01161198.

[15] H. Fastl and E. Zwicker. *Psychoacoustics: Facts and Models*. Springer series in information sciences. Springer Berlin Heidelberg, 2006. ISBN: 9783540231592. URL: https://books.google.com/books?id=0zg9hI586kcC.

[16] Tue Haste Andersen and Kristoffer Jensen. "Importance and representation of phase in the sinusoidal model". In: *AES: Journal of the Audio Engineering Society* 52.11 (2004), pp. 1157–1169. ISSN: 15494950.

[17] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments, 2nd Edition*. New York: Springer Verlag, 1998. ISBN: 0-387-98374-0.

[18] Hanna Järveläinen, V. Välimäki, and M. Karjalainen. "Audibility of inharmonicity in string instrument sounds, and implications to digital sound synthesis". In: Oct. 1999, pp. 359–362.

[19] Karl Kristoffer. "Timbre Models of Musical Sounds Kristoffer Jensen Datalogisk Institut , Københavns Universitet Department of Computer Science , University of Copenhagen". In: September (1999).

[20] Stephen McAdams et al. "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes". In: *Psychological Research* 58.3 (1995), pp. 177–192. ISSN: 03400727. DOI: 10.1007/BF00419633.

[21] Antti Eronen. "Comparison of features for musical instrument recognition". In: *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* October (2001), pp. 19–22. DOI: 10.1109/aspaa.2001.969532.

[22] Vincent Lostanlen and Carmine Emanuele Cella. "Deep convolutional networks on the pitch spiral for music instrument recognition". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016* (2016), pp. 612–618.

[23] Siddharth Gururani, Mohit Sharma, and Alexander Lerch. "An attention mechanism for musical instrument recognition". In: *arXiv preprint arXiv:1907.04294* (2019).

[24] Yun Ning Hung and Yi Hsuan Yang. "Frame-level instrument recognition by timbre and pitch". In: *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018* (2018), pp. 135–142.

[25] *London Phiharmonia sound samples library*. https://philharmonia.co.uk/resources/sound-samples/. Accessed: 2020-11-30.

[26] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *TEST* 25.2 (2016), pp. 197–227. DOI: 10.1007/s11749-016-0481-7. URL: https://doi.org/10.1007/s11749-016-0481-7.