Automatic Estimation of FM Synthesis Parameters by Convolutional Neural Network

Sergio Rocha da Silva
sergio.silva16@unifesp.br
Universidade Federal de São Paulo
São José dos Campos, São Paulo, Brasil

Abstract

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

ACM Reference Format:
Sergio Rocha da Silva. 2026. Automatic Estimation of FM Synthesis Parameters by Convolutional Neural Network. In *Proceedings of The 41st ACM/SIGAPP Symposium on Applied Computing (SAC'26).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

2 Theoretical foundations

2.1 Musical Computation

Musical computation is a broad and interdisciplinary field that encompasses all aspects of musical processing through a computer. It is beyond the scope of this article to provide a comprehensive overview of this field; however, for the purposes of this study, it is important to cover the basic elements of sound and the fundamentals of audio digitization.

2.1.1 Sound elements. Audio is a mechanical compression wave that travels through the air. This type of wave can be converted into an electrical signal through microphones, which consist of an arrangement of coils and magnets. As a wave, its main elements are:

- Pitch: The fundamental frequency of a sound, which determines whether it is low or high. The higher the basis frequency, the higher the pitch.
- Duration: The length of time a sound is sustained and can be heard.
- Loudness (or amplitude): The intensity or energy level of a sound. In terms of a wave, it corresponds to the amplitude of the sound wave.

There are other relevant concepts that could be discussed; however, for the sake of simplicity (and because this work is focuses on Machine Learning), only these will be mentioned.

However, it is important to highlight that although sound can be described in terms of the aspects mentioned above, hearing is mainly a cognitive process. Therefore, sound comparison cannot be fully addressed by mathematical tools, leaving a significant space for perceptual comparison and classification (not to mention aspects related to personal preferences).

2.1.2 Sound timbre. Another important aspect of acoustic sounds is timbre, which can be understood as the "color of sound."

From a mathematical point of view, timbre varies according to the harmonics present in a sound sample, or, in other words, according to the set of oscillations that compose the sound.

In the case of acoustic instruments, what happens is that the air vibrations produced by the instrument are not simple pure sinusoids. In fact, when a string vibrates, or when air resonates inside a tube, many parts of the instrument vibrate simultaneously, and the frequency with which each part vibrates differs according to its composition, position, and other factors.

The result is that each instrument has a unique sound, especially in the case of handcrafted wooden instruments such as violins or cellos.

To obtain a clear visualization of this aspect, it is common to convert a sound sample from the time domain to the frequency domain using a Fourier Transform technique.

For example, consider the two images below, which compare the same violin audio sample in both the time and frequency domains:

s the figures show, there is a main frequency that defines the pitch of the sound, along with many other frequencies that allow the ear to identify the instrument and distinguish between different kinds of sounds.

2.1.3 Audio digitalization. A microphone converts a pressure wave into an electrical waveform. However, in computational music, the audio signal must be discretized to be represented as a digital audio file. The primary process for digital audio representation is sampling.

SAC'26, Thessaloniki, Greece

ACM ISBN 979-X-XXXX-XXXX-X/26/03
https://doi.org/XXXXXXX.XXXXXXX

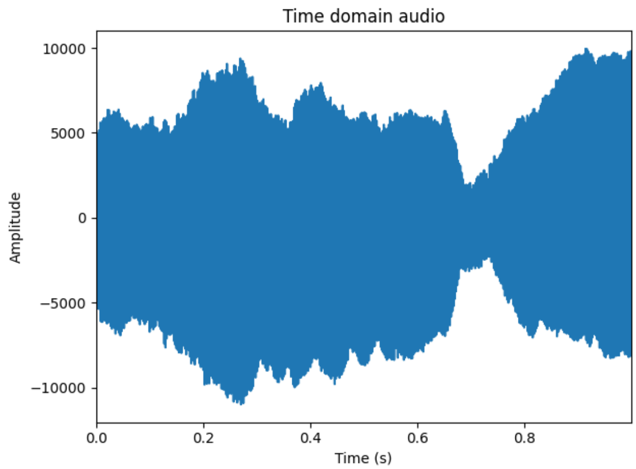

Figure 1: One second time domain violin sample.

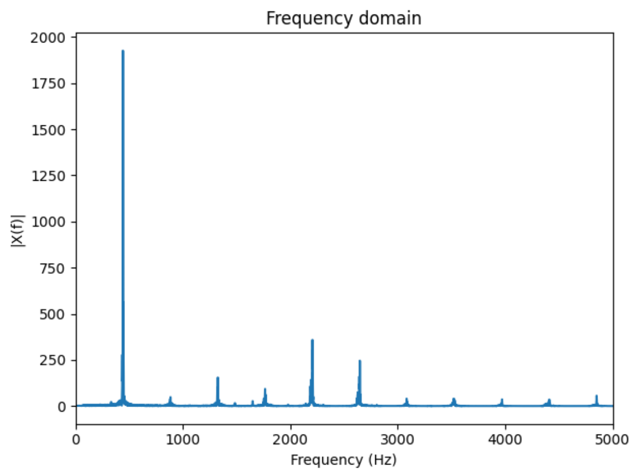

Figure 2: TThe same sample in frequency domain.

Sampling consists of recording a sequence of wave amplitudes at fixed intervals. For example, if a sound is recorded using a sampling rate of 22,100 Hz, it means that the computer will capture 22,100 audio samples per second.

This approach is very efficient because human hearing can detect frequencies between 20 Hz and 20,000 Hz. However, for most musical sounds, the maximum frequency is typically below 10 kHz (in fact, even a violin — one of the highest-pitched instruments — reaches only around 3,500 Hz). And according to the Nyquist-Shannon Theorem, to properly sample a periodic signal, the sampling rate must be at least twice the frequency of the highest component of the signal.

Thus, 22,100 Hz can be a good sampling rate for recording music while saving disk space. However, if the goal is to cover the full range of human hearing, higher rates, such as 44,200 Hz, can also be used.

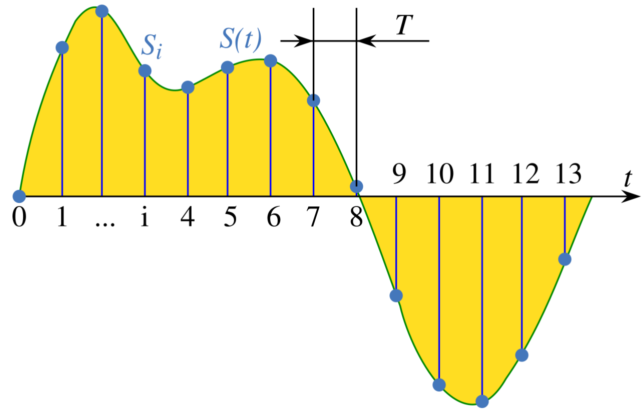

Figure 3: Signal sampling. Source: [1]

Note that with this process, all sound elements highlighted before, can be recorded and reprocedure later.

2.1.4 Sound synthesis. After the development of sound recording techniques, musical computation evolved toward the creation of artificial waveforms. Of course, before digital audio synthesis, there were several techniques that allowed analog synthesis. However, the underlying idea is essentially quite similar.

In summary, audio synthesis techniques can be divided into three basic types:

- Additive synthesis: Perhaps the most straightforward idea from a mathematical point of view. It consists of generating each frequency component of a desired timbre separately and then summing them. In fact, it is more like a weighted sum, since each component is weighted according to its contribution to the timbre. A normalization step is commonly applied after the process to avoid amplitude overload.
- Subtractive synthesis: The opposite of the previous technique, but historically more common due to its effectiveness. In short, it consists of taking a complex waveform, rich in harmonics, and applying various filters to obtain the desired sound. It is more empirical than additive synthesis (which is based on spectral analysis), but simpler to implement and less computationally intensive. Typically, the original waveforms are complex ones such as the sawtooth wave, square wave, or triangular wave.
- FM synthesis: Another empirical approach to sound synthesis, which will be addressed in detail in a later section. In short, it consists of altering a simple waveform by applying a high-frequency modulation, not to the generated waveform itself, but directly to the frequency parameter of the carrier oscillator. This kind of distortion of the base frequency produces mirrored harmonics that enrich the timbre of the sound. It is up to the musician to choose the appropriate parameters to produce the desired timbre.

There are many other techniques that could be mentioned; however, these three represent the most direct forms of synthesis, or the most purely mathematical ones, in the sense that they do not require manipulation of pre-recorded audio.

2.1.5 ADSR Audio envelop. Another important technique for achieving good results is the so-called "Attack, Decay, Sustain, and Release (ADSR) envelope", which consists of applying a multiplicative function to the waveform, altering only its amplitude.

The basic idea is to control the sound intensity by simulating the natural behavior of acoustic sounds. Typically, an ADSR envelope is defined as a function with a domain between 0 and 1, which is applied to the generated waveform:

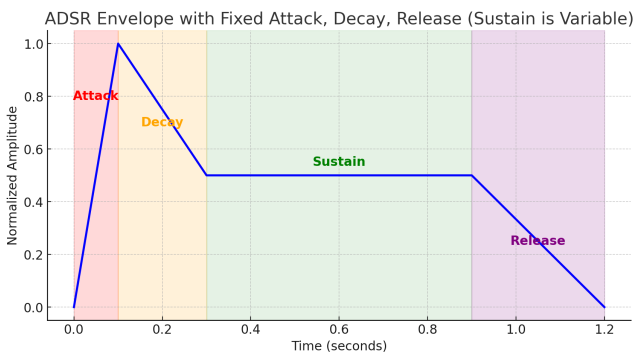

Figure 4: ADSR linear envelop.

The ADSR envelope can be understood as an amplitude function:

$$y(t) = e(t) \cdot s(t)$$

Where:

- $s(t)$: the raw generated sound wave.
- $e(t)$: the ADSR envelope function (normalized amplitude ranging from 0 to 1).
- $y(t)$: the resulting sound.

Although real audio samples can show a wide variation of envelope behaviors, it is common practice to use a linear function to achieve the goal, since the human ear is more sensitive to abrupt changes and to relative differences. However, for longer variations, the ear can still perceive the change, which is why some synthesizers allow the use of exponential or logarithmic functions as well.

2.2 FM Synthesis

As shown in the previous sections, one of the most popular sound synthesis techniques is Frequency Modulation Synthesis (FM Synthesis), which was introduced by John M. Chowning in his famous paper "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation", published at the Stanford Artificial Intelligence Laboratory.

Chowning explored the effects of frequency modulation when applied with high frequencies, close to the audible range, instead of the traditional use of inaudible frequencies.

In fact, the concept of frequency modulation was already well known before this paper; however, it had been used mainly in radio transmission of music or to apply slight distortions through low-frequency oscillators (LFOs), producing effects such as vibrato.

As Chowning himself pointed out in his paper, when the technique is applied to vary the frequency of the carrier wave through a high-frequency modulator wave, it "results in a surprising control of audio spectra" (Chowning).

In summary, considering only two oscillators (the main one, called the carrier, and the modulator) the instantaneous frequency of the output audio can be expressed by the following function:

$$f(t) = f_c + I \cdot \sin(2\pi f_m t)$$

Where:

- $f(t)$: the instantaneous frequency of the generated sound wave.
- f_c: the original frequency of the carrier wave.
- I: the modulation index, which defines the intensity of the modulation (the greater the value, the more harmonics are generated).
- f_m: the frequency of the modulator wave (which determines the spacing of the harmonics).
- t: the time variable.

The most important point, however, is that this simple manipulation results in symmetric sidebands around the carrier frequency, spaced at multiples of f_m.

For a more intuitive understanding, if a carrier frequency f_m is modulated by a frequency f_m, the resulting spectrum will contain:

$$f_c \pm n f_m$$

For example, by choosing $f_c = 440$ and $f_m = 220$, the resulting wave will contain frequency components such as 220 Hz, 440 Hz, 660 Hz, 880 Hz, and so on, with decreasing amplitudes (the amplitudes follow the Bessel coefficients, but this discussion is beyond the scope of this article).

The last important characteristic of Frequency Modulation Synthesis concerns the ratio between the carrier frequency and the modulator frequency f_c/f_m, since this ratio directly influences the perceived audio result.

This ratio determines whether the generated harmonics align with the natural harmonic series (widely used by musicians) or not:

- If the ratio is an integer number, the harmonic sidebands (in the spectral view) will appear as integer multiples of f_c, generating harmonic sounds similar to those of real instruments (consistent with traditional music theory).
- If the ratio is not an integer value, the harmonic sidebands will not align perfectly with f_c, and the generated sound may appear metallic or dissonant (this effect is useful for creating sounds such as bells or for producing special audio effects).

2.3 Audio comparison

The task of audio comparison is not straightforward, especially when the goal is to achieve equality in terms of human perception. It can be complex, computationally demanding, and inherently subjective.

However, as described in the section on timbre, one intuitive and efficient way to approach this problem is by comparing the audio in its spectral representation, since timbre is precisely defined by this combination of frequencies.

That said, this approach alone is not sufficient, because the Fourier Transform considers the entire audio signal at once, mixing different parts of a sound sample. For example, if a sample contains three different pitches in sequence, its frequency-domain representation will show them simultaneously, which may give the impression that the timbre is a mixture of all these pitches (similar to a chord rather than a sequence).

In fact, according to Claesson (2019), FFT-based spectral comparison is valid but only a part of the solution. He highlights the following additional metrics:

- FFT Distance: This consists of computing the Fast Fourier Transform of the two samples being compared and then calculating the normalized Euclidean distance between them (the maximum distance is 1, and the minimum is 0).
- Short Time Fourier Transform (STFT) Distance: Equivalent to the FFT Distance, but calculated over short, overlapping slices of the sample, which provides time-localized spectral information.
- Log-Mel Spectogram Distance: Similar to the previous metrics, but based on the Log-Mel Spectrogram, which is essentially an adaptation of the FFT spectrum that incorporates human auditory perception. In summary, the Log-Mel Spectrogram is obtained by computing the STFT, mapping the frequencies onto the Mel scale (explained later in this document), and finally projecting the amplitudes onto a logarithmic scale.

In addition to these three metrics, Claesson also recommends the use of Euclidean Distance of the Envelope to compare audio envelopes. This approach can be very useful, but since this work is focused on the similarity between original and synthesized audio in terms of timbre, the envelope comparison, while important for real-world applications, falls outside the scope of the present study.

2.3.1 Mel-Scale Representation. The Mel scale was proposed by Stevens, Volkmann, and Newman in 1937 (Claesson, 2019), and its main purpose is to reflect the human perception of frequency.

The key difference from the linear scale is that, as the values on the scale increase, the perceptual difference between two adjacent points becomes smaller (and the corresponding difference in hertz is also reduced).

In fact, the Mel scale maps linear frequencies onto a logarithmic scale. A frequency f s converted to the Mel scale through the following formula:

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

For a better understanding of this conversion, consider the following graph, which compares the linear hertz scale with the Mel scale:

It is a very useful metric for audio comparison, as it takes into account the human auditory perception of frequency.

2.4 Convolutional Neural Networks (CNN)

2.4.1 2D Convolutional Neural Networks. In summary, a convolutional neural network is based on the idea of filters.

Before the introduction of CNNs, was already known that many computer vision problems can be solved through the evaluation

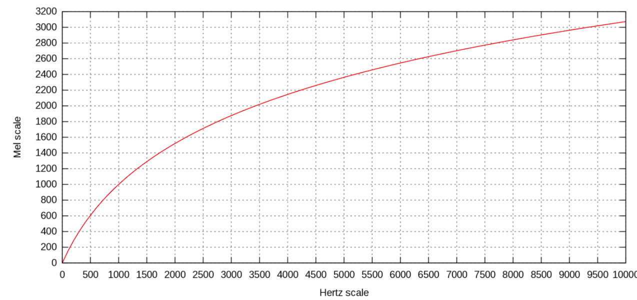

Figure 5: Comparision between linear hertz scale and Mel scale (Claesson, 2019).

of image features, which become more evident depending on the filters applied.

For exampe, to classify differe animals, just the silhouettes can be enough. Of course, the fewer features considered, the higher the chance of mistakes, but, as popular saying goes: “if it walks like a duck, quacks like a duck, flies like a duck… it's a duck!”. That's the basic idea.

So, the CNNs was concept in such a way that several layers of filters of fixed size (chosen by the network designer) are defined, in addition to some pooling layers (which will be discussed later).

These filter layers are also organized according to the designer's choices (usually based on many experiments) and can be applied in sequence, in parallel, using different combinations, and so on.

However, it is important to note that the filters serve as the link between the different representations of the input image built within the network.

That is, although the network still takes an image as input, and although this image is processed to extract the various features needed for the problem, there are no dense interconnections between these image representations.

In fact, what interposes between one image representation and another is precisely a CNN filter. And even this filter is not densely connected to the image representations.

In practice, the filter is not connected to any specific part of the image. Instead, it is “slid” across the image, performing a Hadamard multiplication (element-wise product) between the filter and each frame of the image (of the same size), producing, as output, a “new filtered image” (multiplication, element by element, of two matrices of the same size, with a final sum of all elements — and, typically, with the application of an activation function to the result).

Visually, the work can be represented as follows:

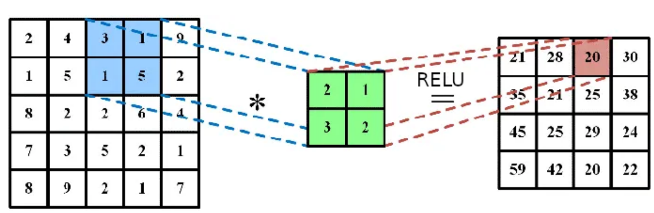

Figure 6: Hadamard product between a filter and an input layer slice (Adapted from ResearchGate, 2024)).

Table 1: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
\$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

And it is precisely from this sliding operation that the network architecture takes its name, since this action corresponds to the mathematical operation of convolution:

To an inout image I, a filter (or Kernel) K, the convolution output O, in a point (i, j) is given by:

$$O(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} K(m,n) \times I(i+m, j+n)$$

Where:

- $K(m, n)$: the filter element on position (m, n).
- $I(i+m, j+n)$: the element of the input on position of current window.
- M, N: the height and width of the filter.
- $O(i, j)$: the output element on position (i, j).

Thus, CNNs can perform image filtering while at the same time reducing the amount of memory required compared to traditional fully connected feed-forward networks (because memory is allocated only for the filters and their results, rather than for each individual interconnection weight).

After extracting the necessary features from the data, a small traditional dense (multi-layer) network is added after the convolution (and pooling) layers. This dense network is responsible for identifying the patterns required to predict the target variable (whether for classification or regression).

A CNN is typically divided into two stages. The first stage consists of the application of filters (along with pooling layers), while the second stage applies a fully connected network. In this second stage, the output of the first part undergoes a flattening process so that it can be used as input to the fully connected layers, which handle the usual tasks of classification or regression.

3 Related work

4 Proposed approach

5 Experiments

5.1 Plan

-dataset -protococlos de avaliação

5.2 Results

6 Conclusion

7 Tables

The "acmart" document class includes the "booktabs" package. Table captions are placed *above* the table.

8 Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

Inline equations: $\lim_{n\rightarrow\infty} x = 0$, set here in in-line math style.

Numbered equation, shown as an inline equation above:

$$\lim_{n\rightarrow\infty} x = 0 \tag{1}$$

Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_0^{\pi+2} f \tag{2}$$

9 Figures

The "`figure`" environment should be used for figures. One or more images can be placed within a figure. If your figure contains third-party material, you must clearly identify it as such, as shown in the example below.

Figure 7: 1907 Franklin Model D roadster. Photograph by Harris & Ewing, Inc. [Public domain], via Wikimedia Commons. (https://goo.gl/VLCRBB).

Your figures should contain a caption which describes the figure to the reader.

Figure captions are placed *below* the figure.

Every figure should also have a figure description unless it is purely decorative. These descriptions convey what's in the image to someone who cannot see it. They are also used by search engine crawlers for indexing images, and when images cannot be loaded.

A figure description must be unformatted plain text less than 2000 characters long (including spaces). **Figure descriptions should not repeat the figure caption – their purpose is to capture important information that is not already provided in the**

caption or the main text of the paper. For figures that convey important and complex new information, a short text description may not be adequate. More complex alternative descriptions can be placed in an appendix and referenced in a short figure description. For example, provide a data table capturing the information in a bar chart, or a structured list representing a graph. For additional information regarding how best to write figure descriptions and why doing this is so important, please see https://www.acm.org/publications/taps/describing-figures/.

Acknowledgments

This work is supported by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), grants 406417/2022-9, 102475/2024-5, and 312209/2022-3, and *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP), grant 2020/09835-1 (CPA IARA).

References

[1] Myriam KRASILCHIK, Celso de Barros Gomes, Francisco Inácio Homem Melo, Heliodoro Teixeira Bastos Filho, Julio Roberto Katinsky, and Tupã Gomes Corrêa. 1996. A USP e sua identidade visual. Tech. rep. São Paulo. http://www.scs.usp.br/identidadevisual/wp-content/uploads/myriamkrasilchik_1996.pdf.