Analyses of Convolutional Neural Networks for Automatic tagging of music tracks

Master Thesis

Aravind Sankaran

Supervisors

Prof. Paolo Bientinesi

Prof. Marco Alunno

Examiners

Prof. Paolo Bientinesi

Prof. Bastian Leibe

Abstract

We address the automatic music tagging problem that can be solved by training with a personal repertoire. Trying to convey the meaning of music or its emotional content is not an easy task. This leaves a semantic gap between music audio and listener's choice of words to describe the aspects of music. Furthermore, some emotional and structural components of music are realized only after listening to a greater length of the song. Hence we address the problem of multi-label classification on features that approximates the whole song. Since personal repertoires are usually small, we stick to find solutions on a medium sized dataset. We use convolutional neural networks (CNNs) for localized feature extraction. These features are extracted every 29s from log-amplitude Mel-spectrogram and fed into sequence-to-one recurrent neural network for temporal summarization of the song, which are then used for classification. We finetune over CNN architectures in [8] [9] and report the categorical AUC scores and Mean average precision. We also report the variations in performance with increasing number of training labels. [TODO : findings]

Acknowledgments

Contents

Chapter 1

Introduction

Music has its own language and describing it can be quite tricky. Talking about music may simply require as much vocabulary as any technical subject, but a strict vocabulary signal mapping can steal the artistic freedom of expression. Musicians and composers usually discuss their work with jargons describing a particular note, chord, sound, or rhythm and how the pieces are put together as a whole. At the same time, with the amount to music recordings constantly growing, it would be time consuming for anyone to manually tag every one of them. Hence the ability to automatically summarize such tags according to personal or specific glossary is studied. In section 1, the motivation for research is explained by describing some shortcomings of the current state of art.

1.1 Motivation

Automatic music recommendation has become an active area of research in recent years because a lot of music is now sold and consumed digitally. These recommendation algorithms allows listeners to discover the songs that match their taste. It also enables online music stores to filter their target audience. However, current state of art methods still suffer from the following problems mentioned below.

1.1.1 Cold start problem with collaborative filtering methods

In the area of music information retrieval, great technical progress have been made to enable efficient retrieval and organization of audio content. But the problem of music recommendation is however complicated because of the sheer variety of genre,mood,acoustic scene, as well as social factors that affect listeners preference. When the usage data is available, one can use collaborative filtering to recommend the tracks on trending lists. In absence of such usage data, one resorts to content-based methods, where just the audio signal is used for generating recommendations. Although collaborative filtering techniques have shown to outperform content-based recommendations [5], they suffer from cold start problem, making it less efficient for new and unpopular songs.

1.1.2 Problems with content based methods

Content-based recommendation methods map audio signal to acoustic cues, which are then used for retrieval.

Psychoacoustic assumptions

The aspect of music that grabs attention is still an ongoing research in experimental psychology. Aesthetic judgments are strongly (although not exclusively) influenced by cultural variables, although there may be some universals that set constraints on what we initially find beautiful[no accounting for taste]. Hence, using currently available datasets[4][3] for training can make recommendation systems suffer from generalization assumptions.

Temporal summarization of audio content

The current state of art music tagging algorithms[9][6] are established by training on datasets that contain just short music excerpts. But choice of music is usually done based on the aspects of entire song. The ability to summarize the aesthetic judgement based on sequence of section-wise tags have not yet been studied. Furthermore, there arent much public datasets that summarize tags for entire song.

1.1.3 Need for adaptive glossary

Current recommendation systems including the ones that use collaborative filtering, restrict the user with the choice of tags. Moreover, it is not guaranteed that all users will perceive all the tags in the same way. A recent study in idiographic music psychology have indicated that different people use different aesthetic criteria to make judgement about music[10]. Hence there is a need to study the performance of recommendation algorithms trained on personal repertoire..

1.2 Structure of the Thesis

In the following work, only content-based methods are considered for multi-label classification. That is to say that only raw signal is used as input for classification. This requires feature extraction from input audio, temporal summarization of features followed by classification.

1.2.1 Convolutional neural network for feature extraction

The input signal preprocessed to spectrogram is mapped to the feature space by convolving hierarchically with learnable filters. (see ch.2,1) Conceptual arguments have been made to demonstrate that deep processing models are powerful extensions of hand-crafted feature extraction methods[7]. (see ch.3.1) It is also shown that deep layers learn to capture textures and patterns of continuous distribution on a spectrogram for music classification task. [explaining cNN for music class]. (see ch 3.2). we discuss the ability of these CNN models to be fine tuned on a medium sized dataset (see ch). Convolutions over log amplitude mel spectrogram and MFCC are studied (see ch)

1.2.2 Recurrent neural network for temporal summarization of features

The features extracted on every 29.1s time frame are then sent to sequence to one RNN. This leaves us with a feature of fixed dimension for audio of arbitrary length up to five minutes. (see ch). Here we make an assumption that a listener can make an aesthetic judgement within 5 minutes. Conceptual comparisons of RNN with the state of art temporal feature pooling technique in [ ] have been discussed (see) . Effectiveness of temporal summarization have been justified by comparing with the performance of section-wise merging of tags (see)

1.2.3 classification

The features are then mapped to the probability space of labels. Multilayer perceptron with binary cross entropy loss is used for training. End to end training is compared with two-stage method, separating training of features and training of classifier (see ch). In the two-stage approach, MLP is compared with SVM classifier. (see)

1.2.4 supervised training on repository with aesthetic tags

A properly labelled training set is usually required to solve the task of automatic tagging. In this thesis, a repository labelled with aesthetic judgements of songs is used. An aesthetic judgement is a subjective evaluation of a piece of music as art based on an individual set of aesthetic properties. [from everyday ..] An aesthetic judgement is assumed to rely more on higher cognitive functions , domain relevant knowledge, and a fluid, individualized process that may change across time and context. We discuss how aesthetic judgements influence preference for a song. In a recent work in experimental music psychology, strong correlation have been found between the two. But it is also shown that such preferences also vary widely between cultural contexts. (see) This also justifies the need for adaptive glossaries in recommendation systems.

1.3 Outline of the report

In chapter 2, the terminologies and mathematical formulations are elaborated. Advanced readers can skip this chapter. In chapter 3, a detailed overview of previous research, their shortcomings for the current problem along with justification for proposed models are discussed. In Chapter 4, details of the dataset, implementations and the experiment results of proposed models are discussed. In chapter 5, the results are analysed and the need for biologically motivated feature extraction techniques are discussed.

Chapter 2

Fundamentals

In this chapter, acoustical characteristics of music signal that enables general MIR tasks will be introduced.We will examine the Fourier Series representations of sound waves and see how they relate to harmonics and tonal color of instruments

2.1 Representation of audio signal

The traditional way of observing signals is to view them in the time domain. The time domain is a record of what happened to a parameter of the system versus time. Standard formats use change of amplitude with time. However, it is useful to change the representation to frequency domain of the signal, which is also called spectrum. This is simply because our ear-brain combination is an excellent frequency domain analyzer. The ear-brain splits the audio spectrum into many narrow bands and determines the power present in each band. Hence, it can easily pick small sounds out of loud background noise. [pp1]

It was shown over one hundred years ago by Baron Jean Baptiste Fourier that any waveform that exists in the real world can be generated by series of sinusoids which are a function of frequencies. Hence any stationary signal (i.e signal at time t) can be represented as a function of a fundamental frequency (lowest frequency), and other frequencies which are multiples of fundamental. The following abstract representation is adapted for further explanations in this section

EQ

2.1.1 Discriminants of music signal - Harmonics and overtones

When a note is played on an instrument, listeners hear the played tone as the fundamental, as well as a combination of its harmonics sounding at the same time (pitch) (Hammond, 2011). Harmonics are tones that have frequencies that are integer multiples of the fundamental frequency. The fundamental and its harmonics naturally sound good together.

EQ, ki = integers

These additional frequencies determines the timber of the instrument. The strength, or amplitude, of each harmonic is the difference were hearing, since each note played includes the funda-

mental tone and some harmonics. The instruments timbre is what distinguishes its sound from that of a different instrument.

(graph example of harmonics of two instruments)

The presence of multiple, simultaneous notes in polyphonic music renders accurate pitch tracking very difficult. However, there are many other applications, including chord recognition and music matching, that do not require explicit detection of pitches, and for these tasks several representations of the pitch and harmonic information commonly appear. Usually, there are more instruments being played simultaneously and sometimes accompanied by voices. In such cases, we hear the fundamental and overtones (chord). The overtones are any frequency above the fundamental frequency. The overtones may or may not be harmonics. So overtones are those frequencies which are not just restricted to integer multiples of fundamental. The fundamental and overtones together are called partials

EQ, ki != integers

m1(t) = f(440 ,880)

Where 440 = fundamental, 880 = first harmonic m2(t) = f(330 , 660)

Where 330 = fundamental, 660 = first harmonic

Thus the recorded signal has components of all these frequencies

m(t) = f(330,440,660,880)

Where 330 = fundamental, 440 = second partial, 660 = third partial, 880 = fourth partial

Most certainly, the signal evolves over time and hence the components of frequencies and its amplitudes will vary for each time t. The heat map representation of amplitudes, with frequency along y, time along x is called spectrogram.

Thus to discriminate a signal, we not only need the evolution of frequencies, but also information about harmonics. For instance, to discriminate the instruments from the recorded signal m(t), the classifier should infer the frequencies in the each harmonics m1(t) and m2(t). To discriminate the temporal pattern (Rhythm), we need the evolution of m1 and m2. To identify other aesthetics (warm, city) , the interaction between m1(t) and m2(t) should be inferred. To discriminate voices and other non-harmonic aspects (tempo, beat), the envelop curve of the spectrum will also be needed.

(diagram)

2.1.2 Sampling of continuous-time signal

The digital formats contain the discrete version of the signal obtained by sampling continuous-time signal. For functions that vary with time, let s(t) be a continuous function (or "signal") to be sampled, and let sampling be performed by measuring the value of the continuous function every T seconds, which is called the sampling interval or the sampling period.[1][pp2]. The sampling frequency or sampling rate, fs, is the average number of samples obtained in one second (samples per second),

thus fs = 1/T.

The optimum sampling rate is given by Nyquist-Shannon sampling theorem which says, the sampling frequency (fs) should be at least twice the highest frequency contained in the signal [pp2] Given the human hearing range lies between 20Hz - 20KHz [pp3], most of the digital audio formats

use a standard sampling frequency of 44.4Khz. The signal is further down sampled depending on the kind of feature information needed for classification.

2.1.3 Time-Frequency transformations

The signal represented in the time domain is a set of ordered n-tuples of real numbers $(a_1, a_2, ..., a_N) \in \mathbb{R}^N$ in the vector space V , specifically *Euclidean n-space*. That is to say, a discrete-time signal can be represented as a *linear combination* of Cartesian basis vectors.

$$\mathbf{a}(t) = (a_1, a_2, ..., a_N) = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + ... + a_N\mathbf{e}_N = \sum_{i=1}^{N} a_i\mathbf{e}_i \tag{2.1}$$

where:

$\mathbf{a}$ is a discrete-time signal

$\mathbf{e}_1 ... \mathbf{e}_N$ are Cartesian basis vectors (Unit vectors).

Mapping from time-domain to frequency-domain is looked up on as change of basis. We need to find a set of basis vectors $\boldsymbol{\phi}_\omega$, whose coefficients c_ω then represents the components in frequency domain.

$$\mathbf{a}(t) = \sum_{\omega=0}^{M-1} c_\omega \boldsymbol{\phi}_\omega(t) \tag{2.2}$$

for some integer $0 < M < \infty$. Then $\mathbf{c}(\phi) = (c_0, c_1, ..., c_{M-1}) \in \mathbb{C}^M$ represents the components in frequency domain. Thus our aim is to compute $\mathbf{c}(\phi)$. Computing the Fourier coefficients for periodic and aperiodic signals are discussed below.

Periodic Signals

If $\mathbf{a}(t)$ is periodic in $\mathbf{T}$, then we can apply the definition of **Exponential Fourier Series** expansion and define $\boldsymbol{\phi}$ in equation (2.2) as (See Appendix ??),

$$\boldsymbol{\phi}_k(t) = \frac{1}{\sqrt{T}} e^{ik\omega t} \tag{2.3}$$

Whose basis functions $\boldsymbol{\phi}$ now form *complete orthonormal* set [2]. That is,

$$< \boldsymbol{\phi}_i, \boldsymbol{\phi}_j >= \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \tag{2.4}$$

The fourier series finds a set of discrete coefficients of **harmonically related frequencies** $(k\omega)$. To retrieve c_k, multiply $\boldsymbol{\phi}_k$ on both sides of equation (2.2) and apply the conditions of orthonormality in equation (2.4). Thus

$$c_k =< \mathbf{a}(t), \boldsymbol{\phi}_k(t) > \tag{2.5}$$

Although periodicity assumptions are not made for general music signals, it becomes relevant to deduce rhythmic patterns.

Aperiodic Signals

It is difficult to assume periodicity for a generalized signal. We need to estimate the coefficients $\mathbf{c}$ for continuous frequency variable $\boldsymbol{\omega}$ instead of discrete harmonics $\mathbf{k\omega}$. The Fourier series can not be applied directly and hence Fourier Transform was developed. Here we aim to find out quantity of each sinusoids is the signal $\mathbf{a}(t)$. This can be done by dividing $\mathbf{a}(t)$ by $e^{i\omega t}$ over the time domain. We use the complex exponential in place of sinusoids because we know (see Appendix ??)

$$sin(\omega t + \Phi) \propto e^{i\omega t} \tag{2.6}$$

Where Φ is the phase difference. Thus, the coefficients in the frequency domain are

$$c_\omega = \sum_{t=0}^{N-1} a(t) e^{-i\omega t} \tag{2.7}$$

This is the N-point **Discrete Fourier Transform**. For the proof of existence of such coefficients, please refer to chapter ?? in [2]. From here, $\boldsymbol{\phi}_\omega(t)$ in equation (2.2) can be defined as

$$\boldsymbol{\phi}_\omega(t) = e^{i\omega t} \tag{2.8}$$

Thus, we can compute $\mathbf{a}(t)$ as a linear combination of complex exponentials. This is also known as **Inverse Fourier Transform**.

$$\mathbf{a}(t) = \sum_{\omega=0}^{M-1} c_\omega e^{i\omega t} \tag{2.9}$$

Hence, with Fourier Transform, we can go back and forth between time and frequency domain. It is important to note that these basis vectors need **not** be *orthogonal*.

Fast Fourier Transform(FFT) is an efficient implementation of Discrete Fourier Transform(DFT) which exploits the symmetry of *sines* and *cosines*. While DFT requires $O(N^2)$ operations, FFT requires only $O(NlogN)$ [2].

2.1.4 STFT, Mel-Spectrogram, Chromogram

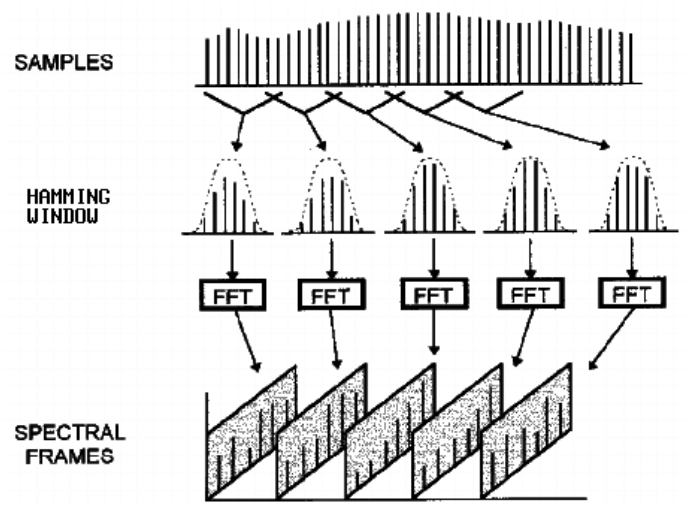
It is useful to perform FFT locally over short segments. This is simply because FFT becomes very expensive for larger N.

$$KNlog(N) < (KN)log(KN)$$

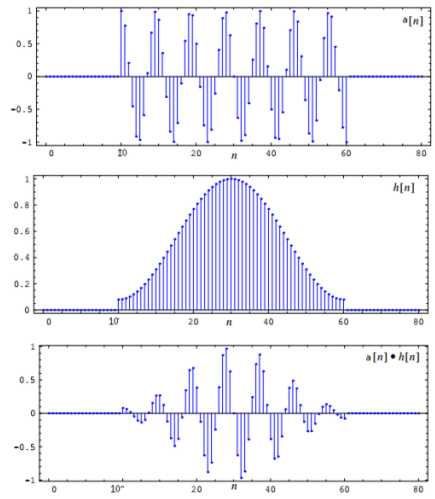
The full length signal is divided into short segments, and FFT is computed separately for each segment. This is known as **Short Time Fourier Transform (STFT)**. Usually the dimension of the frequency components are reduced by using bins. Every frequency component is assigned to it's nearest bin. This however causes **spectral leakage** when we divide the signal into rectangular windows. That is, components at the end of the segment can leak to the adjacent segment. This is avoided by modifying the original signal by applying some window function. The most common window function is the **Hamming Window** defined as,

$$h[n] = 0.54 - 0.46cos(\frac{2\pi n}{N-1}) \tag{2.10}$$

Where $n \in 0, 1, ..., N-1$. The signal approaches zero near $n = 0$ and $n = N-1$, but reaches peak near $n = N/2$ [11]. To overcome the information loss at the ends of the window, signal is divided into segments that are partly *overlapping* with eachother. Figure (2.1 (a)) shows the extraction of spectral frames of a spectrogram.



(a) Windowing is applied on overlapping segments followed by FFT



(b) Application of Hamming Window on a segment of input signal

Figure 2.1: (a) Shows STFT Pipeline. (b) Shows the application of Window function

The discrete STFT (*slow*) for p^{th} frame of signal $\mathbf{a}$ is obtained as,

$$\mathbf{C}(p, \omega) = \sum_{n=p.s}^{p.s+F} \mathbf{a}(n)\mathbf{h}(n - p.s)e^{-i\omega(n-p.s)} \tag{2.11}$$

Where:

P: is the number of spectral frames; $p \in [0, 1..P-1]$

M: is the dimension of discrete frequency space ; $\omega \in \mathbb{R}^M$

F: is the frame length

s: is stride (or) hop-length for the next segment

$\mathbf{C}$: is Fourier Coefficient Matrix ; $\mathbf{C} \in \mathbb{C}^{M \times P}$

$\mathbf{a} \in \mathbb{R}^N$; $n \in [0, 1...N-1]$

$\mathbf{h} \in \mathbb{R}^F$

Equation (2.11) can be seen as a **convolution** over the signal $\mathbf{a}$ with $\mathbf{W}$ which has finite support over the set $\{0, 1.., F\}$ (more details in section ??)

$$\boxed{\mathbf{C}(p, \omega) = \mathbf{a}(n) \star \mathbf{W}_\omega(n - \tau)} \tag{2.12}$$

Where:

$$\tau = p.s$$
$$\mathbf{W}_\omega(n-\tau) = \mathbf{h}(n-\tau)e^{-i\omega(n-\tau)}$$

It is important to note that the coefficients c_ω may be complex valued. They are functions of the amplitude of corresponding sinusoidal component (see Appendix ??). But, to obtain useful metrics, we need to extract some physical quantity from the coefficients. This is where **Parseval's theorem** is used, which relates and time and frequency domain components in DFT as follows [2] :

$$\|\mathbf{c}\|^2 \propto \|\mathbf{a}\|^2 \tag{2.13}$$

If **a** represents amplitude in the time-domain, then as a consequence of Hook's law on energy equation (see Appendix ??), we know that

$$Energy \propto amplitude^2 \tag{2.14}$$

Relating equation 2.11 and 2.12, it can be inferred that **square** of the Fourier coefficients is proportional to the energy distributed in the corresponding frequencies. This is called the **Power Spectrum (E)**. It is often motivating to use this representation because *loudness* is proportional to *energy*.

$$\mathbf{E} = \mathbf{C} \odot \mathbf{C} \tag{2.15}$$

As mentioned earlier, the frequencies in the considered range are grouped into bins. It is useful to do so, not only to reduce dimension but also due to the aliasing effect of human auditory system. This is motivated by the human cochlea (an organ in the ear) which vibrates at different spots depending on the frequency of the incoming sounds. Depending on how frequencies are grouped, two different class of spectrograms are discussed.

Mel Spectrogram

The *mel-scale* was developed to express measured frequency in terms of psychological metrics (i.e perceived pitch). The mel-scale was developed by experimenting with the human ears interpretation of a pitch. The experiment showed that the pitch is linearly perceived in the frequency range 0-1000 Hz. Above 1000 Hz, the scale becomes logarithmic. There are several formulae to convert Hertz to mel. A popularly used formula is noted in [1]

$$\omega_m = 2595 log_{10}\left(1 + \frac{\omega}{700}\right) \tag{2.16}$$

Where ω is the frequency in Hertz. In a mel spectrogram, the frequencies and converted to mels and then grouped into mel-spaced bins. This is done by multiplying the spectrum with some **filter bank (Mf)**. For details about computation of mel-filter banks, refer **??**. Each filter bank is centered at a specific frequency. Hence, to compute R mel bins, we need R mel-filter banks.

$$\mathbf{Mel}(p, \omega_m) = \sum_{\omega=0}^{M} \mathbf{Y}(p, \omega)\mathbf{Mf}(\omega_m, \omega) \tag{2.17}$$

Where:

$\mathbf{Y} = f(\mathbf{C})$

ω_m = mel frequency

When the function f is an defined by equation (2.15), we get **mel power spectrogram**

We can re-write equation (2.17) as,

$$\mathbf{Mel}(p, \omega_m) = \sum_{k=p.M}^{p.M+K} \mathbf{U}(k)\mathbf{M}_{\omega_m}(k - p.M) \tag{2.18}$$

Where:

$K = M.P$ and $k \in [0, 1...K]$

$\mathbf{U}(k) = \mathbf{Y}(i, j)$; $i = floor(\frac{k}{M})$; $j = k - floor(\frac{Mk}{M-1})$

$\omega = k - p.M$

Hence, we can represent mel-spectrogram as **M-strided convolution** over *flattened* $\mathbf{Y}$ with mel filters $\mathbf{M}_{\omega_m}$,

$$\boxed{\mathbf{Mel}(p, \omega_m) = \mathbf{U}(k) \star \mathbf{M}_{\omega_m}(k - p.M)} \tag{2.19}$$

Chromagram

This representation takes advantage of the periodic perception of pitch. Two pitches are perceived similar in "color" if they differ by one or several octaves apart. Chromagram groups such periodic perceptions into same coefficient (chroma). All pitches that belong to the same chroma are said to be from same pitch class. [TODO: How is this computed?]

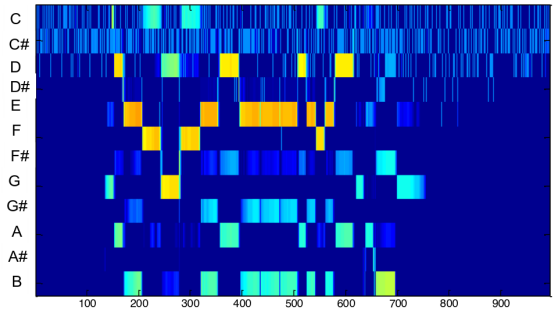


Figure 2.2: Chromagram of Western Pitch Scale

Bibliography

Proceedings

[4] Thierry Bertin-mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. "The million song dataset". In: *In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR.* 2011.

[6] Sander Dieleman and Benjamin Schrauwen. "Multiscale Approaches To Music Audio Feature Learning". In: *ISMIR.* 2013.

[8] Keunwoo Choi, George Fazekas, and Mark Sandler. "Automatic tagging using deep convolutional neural networks". In: *International Society of Music Information Retrieval Conference. ISMIR.* 2016.

Articles

[3] Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Stephen Downie. "Evaluation of algorithms using games: The case of music tagging". In: (2009), pp. 387–392.

[5] M. Slaney. "Web-Scale Multimedia Analysis: Does Content Matter?" In: *IEEE MultiMedia* 18.2 (2011), pp. 12–15. ISSN: 1070-986X. DOI: 10.1109/MMUL.2011.34.

[7] Humphrey Eric J., Juan P. Bello, and LeCun Yann. "Feature learning and deep architectures: new directions for music informatics". In: *Journal of Intelligent Information Systems* 41.3 (2013), pp. 461–481. ISSN: 1573-7675. DOI: 10.1007/s10844-013-0248-5.

[10] Patrik N. Juslin, Laura S. Sakka, Gonalo T. Barradas, and Simon Liljestrm. "No Accounting for Taste? Idiographic Models of Aesthetic Judgmentin Music". In: *Psychology of Aesthetics, Creativity, and the Arts* 10.2 (2016), pp. 157–170. DOI: 10.1037/aca0000034.

Pre-Prints

[9] Keunwoo Choi, Gyorgy Fazekas, Mark Sandler, and Kyunghyun Cho. *Convolutional Recurrent Neural Networks for Music Classification.* Version 3. Dec. 21, 2016. arXiv: 1609.04243.

Books

[1] D. O'Shaughnessy. *Speech communication: human and machine*. Addison-Wesley series in electrical engineering. Addison-Wesley Pub. Co., 1987. ISBN: 9780201165203. URL: https://books.google.de/books?id=mHFQAAAAMAAJ.

[2] R.L. Allen and D. Mills. *Signal Analysis: Time, Frequency, Scale, and Structure*. Wiley, 2004. ISBN: 9780471660361.

Misc

[11] Lecture Notes. *Spectral Leakage and Windowing*. https://mil.ufl.edu/nechyba/www/__eel3135.s2003/lectures/lecture19/spectral_leakage.pdf. Online; accessed 20 March 2017.