UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

Gustavo da Silva Tafarello Salessi

Rotulação Automática de Músicas Usando Redes Neurais Profundas

Campinas
21/08/2020

Gustavo da Silva Tafarello Salessi

Rotulação Automática de Músicas Usando Redes Neurais Profundas*

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. Dr. Marcos Eduardo Ribeiro do Valle Mesquita.

*Este trabalho foi financiado pelo CNPq.

1 Introdução

Rótulos de música (do inglês *music tags*) são metadados atribuídas à um sinal de áudio que transmitem informações de alto nível como ânimo (feliz, triste, raivoso), gênero (jazz, clássica) e instrumentação (violão, cordas, vocal, instrumental) (Choi et al., 2018). Rótulos de músicas são usados, por exemplo, para organização de coleções musicais ou em sistemas de recomendação em serviços de *streaming* de música como *Last.fm*, *Spotify* e *Deezer*. Em geral, os rótulos referem-se tanto à propriedades objetivas como os instrumentos e tipos de vocais (masculino ou feminino) como também propriedades subjetivas como genero (rock, jazz, etc) e ânimo (feliz, triste, etc).

Nesse trabalho, assumiremos que os rótulos são atribuídos por ouvintes ou comunidades em serviços de *streaming*. Rótulos atribuídos por ouvintes, chamados *folksonomias*, geralmente não possuem restrições de quantidade ou vocabulário (Lorince et al., 2015; Lamere, 2008). O *Last.fm* é um exemplo de serviço que usa *folksonomia* para rotular faixas e artistas. Apesar da *folksonomia* não exigir o conhecimento de um especialista, músicas novas ou raramente ouvidas podem não ter rótulos suficientes para suportá-las e, consequentemente, podem gerar problemas no gerenciamento de informações musicais. Por exemplo, músicas não rotuladas ou com rótulos insuficientes podem não ser identificadas num sistema de busca ou não ser recomendadas para potenciais ouvintes (Lamere, 2008). O problema com músicas não rotuladas ou com rótulos insuficientes pode ser superado usando sistemas automáticos para rotulação de músicas (do inglês *music-auto tagging*) (Eck et al., 2008).

Um sistema para rotulação automática de música estabelece uma aplicação entre o conteúdo do áudio e seus rótulos (Choi et al., 2018; Lee and Nam, 2017; Shen et al., 2010). Formalmente, a rotulação de músicas é um problema de classificação multirótulo pois uma faixa musical pode ser associada a mais de um rótulo. Por exemplo, podem ser atribuídos os rótulos instrumental, clássico e alegre a uma faixa de música. Tal como na literatura atual (Lee and Nam, 2017; Choi et al., 2018), nesse trabalho a rotulação automática será realizada por uma rede neural profunda aplicada no espectrograma na escala mel (do inglês *mel-spectrogram*). Sobretudo, o treinamento da rede será realizado considerando um conjunto significativo de dados de treinamento obtidos usando *folksonomia* (Bertin-Mathieux et al., 2011). Com efeito, além de ser compacto, acredita-se que o espectrograma na escala mel fornece características suficientes

da música para o problema de rotulação (Choi et al., 2018). Redes neurais profundas apresentaram excelentes resultados em problemas de classificação (Choi et al., 2016), reconhecimento de padrões (Wu et al., 2018), sistemas de recomendação (Zhang et al., 2019), análise de imagens médicas (Ker et al., 2017) e processamento de linguagem (Merity et al., 2018). Elas também foram efetivamente aplicadas em problemas de rotulação de músicas. Sobretudo, estudos recentes mostram que certas redes neurais profundas são robustas no problema de rotulação de música mesmo quando o conjunto de treinamento é fraco (isto é, ruidoso, incompleto ou inconsistente), como é o caso dos dados de *folksonomia*.

2 Conceitos de Sinais Contínuos e Discretos no Tempo

O ponto inicial é entender o conceito de "sinal". De uma maneira fundamental, a palavra sinal se refere ao processo de transmitir informações em algum formato. Um sinal pode ser representado de diversas maneiras, porém a informação sempre está contida em algum tipo de variação (Oppenheim and Willsky, 2010). De forma matemática, sinais podem ser representados como funções de uma ou mais variáveis independentes. A Figura 1 mostra um sinal representado por uma função de uma única variável (tempo), o foco de estudo desse projeto. O som consiste da variação da pressão do ar no tempo, portanto é um sinal, daí a importância de se estudar esse conceito neste projeto. Dentro do computador o sinal de som é representado como um sinal digital.

2.1 Sinais de Tempo Contínuos e Discretos

Dois tipos muito usados de sinais são os de tempo contínuo e os de tempo discreto. No caso do primeiro a variável independente de tempo é contínua e portanto esses sinais são definidos em um conjunto contínuo, ou seja, sua variação se dá de forma gradual conforme ilustra a Figura 2. Sinais de tempo discreto são definidos somente em instantes discretos de tempo, ou seja, a sua variação se dá de forma abrupta conforme mostrado na Figura 3.

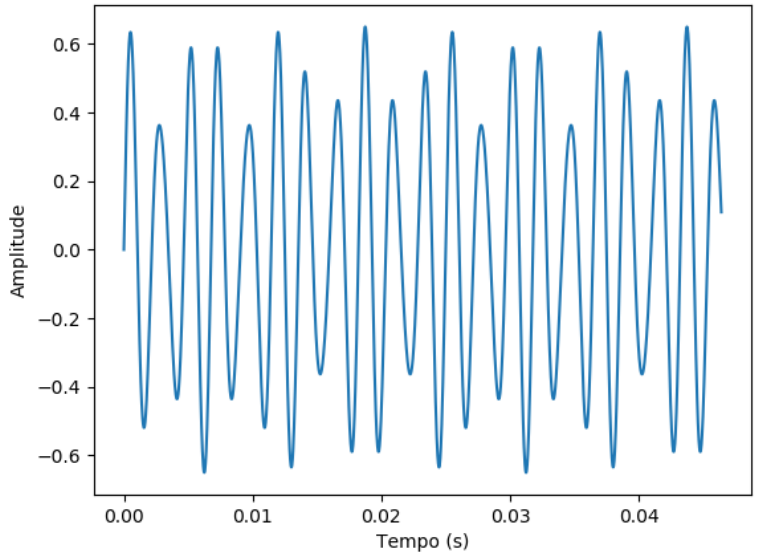

Figura 1: Gráfico do sinal $h(t)$.

3 Transformada de Fourier

Uma técnica de se resolver problemas muito complexos é dividi-los em problemas menores. Os problemas menores são resolvidos e combinados para formar a solução do problema original (Phillips and Parr, 1995). Outra ferramenta importante é salientar as características fundamentais de um determinado sinal (Cadzow and Landingham, 1985). São nesses pontos que a série de Fourier e a transformada de Fourier se tornam importantes ferramentas na análise de sinais.

3.1 Série de Fourier

Uma função $f(t)$, com $t \in \mathbb{R}$, pode ser representada como uma soma de senos e cossenos, da seguinte forma (Vaz Jr. and de Oliveira, 2016):

$$f(t) = \frac{ao}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \operatorname{sen} nt). \tag{1}$$

Essa é a representação em série de Fourier (SF) da função $f(t)$. Ela descreve a função apenas no intervalo $-\pi \leq t \leq \pi$, para obter uma representação em um intervalo arbitrário L, basta

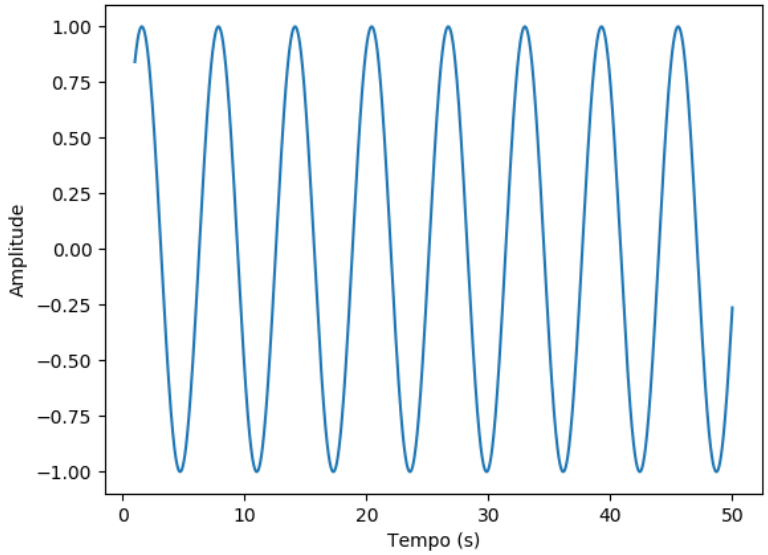

Figura 2: Exemplo de um sinal de tempo contínuo

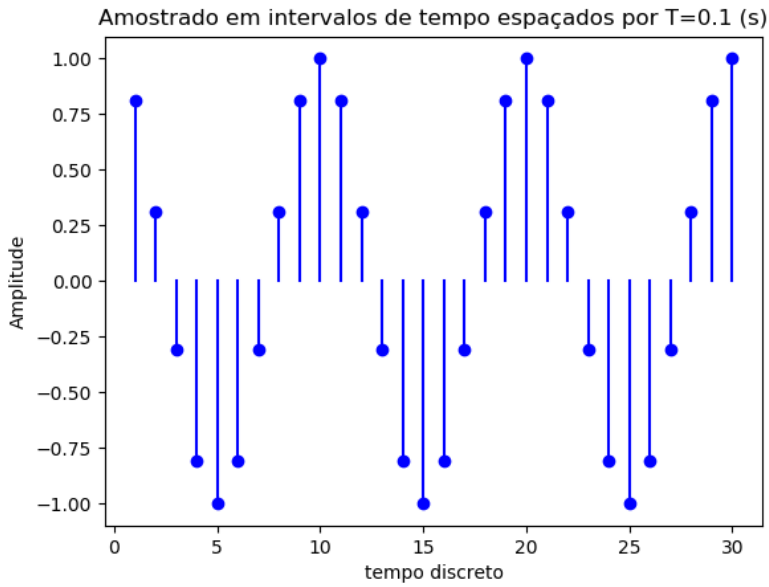

Figura 3: Exemplo de um sinal de tempo discreto

fazer uma mudança de variável e a SF se torna (Vaz Jr. and de Oliveira, 2016):

$$f(t) = \frac{ao}{2} + \sum_{n=1}^{\infty}(a_n \cos\frac{n\pi t}{L} + b_n \operatorname{sen}\frac{n\pi t}{L}). \tag{2}$$

O sinal pode então ser representado como uma combinação linear de um número infinito de senos e cossenos. Fazendo a decomposição de um sinal em série de Fourier ele está sendo expresso em uma soma ponderada infinita de sinais elementares, que são muito mais simples de usar e analisar.

A série de Fourier pode ser representada na sua forma complexa, para isso se utiliza a fórmula de Euler:

$$e^{it} = \cos t + i \operatorname{sen} t. \tag{3}$$

A série de Fourier para um intervalo arbitrário L fica

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{\frac{in\pi t}{L}}. \tag{4}$$

Nota-se que agora a série começa em $n = -\infty$, pois definiu-se c_{-n} igual ao conjugado complexo de c_n, ou seja, $c_{-n} = {c_n}^*$. O coeficiente c_n é determinado pela equação:

$$c_n = \frac{1}{2L}\int_{-L}^{L} f(t) e^{\frac{-in\pi t}{L}} dt. \tag{5}$$

3.2 Transformada de Fourier de Tempo Contínuo

A transformada de Fourier (TFC) de uma função $f(t)$ é definida segundo a equação (Phillips and Parr, 1995):

$$\mathcal{F}(p) = \int_{-\infty}^{+\infty} f(t) e^{i(2\pi pt)} dt, \qquad \forall p \in \mathbb{R}. \tag{6}$$

A TFC transforma a função $f(t)$ do domínio de tempo para o domínio das frequências. Com efeito a transformada expressa a função como uma combinação linear de senos e cossenos sobre todas as frequências possíveis.

3.3 Transformada de Fourier de Tempo Discreto

Considerando o caso em que se tem apenas um número finito de valores de $f(t)$ para o intervalo finito $[0, T]$, o espaço é dividido em N pontos t_k dados por.

$$t_k = k\frac{T}{N}, \qquad \forall k = 0, 1, ..., N-1. \tag{7}$$

A transformada de Fourier da função $f(t)$ em tempo discreto é dada pela equação (Vaz Jr. and de Oliveira, 2016):

$$\mathcal{F}(n) = \sum_{k=0}^{N-1} f(t_k) W^{nt_k}, \qquad \forall n = 0, 1, ..., N-1. \tag{8}$$

em que $W = e^{i2\pi/T}$

3.4 Transformada Rápida de Fourier

Utilizando a transformada de Fourier de tempo discreto dada por (8), e substituindo t_k/T por k/N, pode-se então separar a transformada em duas partes:

$$\mathcal{F}(n) = \sum_{k=0}^{(N/2)-1} f(t_k) e^{i(2\pi nk)/N} + \sum_{k=(N/2)}^{N-1} f(t_k) e^{i(2\pi nk)/N}. \tag{9}$$

Introduzindo $p = k - (N/2)$ no segundo somatório em (9) tem-se:

$$\mathcal{F}(n) = \sum_{k=0}^{(N/2)-1} f(t_k) e^{i(2\pi nk)/N} + \sum_{p=0}^{(N/2)-1} f(t_{p+(N/2)}) e^{i(2\pi n(p+(N/2)))/N},$$

ou ainda

$$\mathcal{F}(n) = \sum_{k=0}^{(N/2)-1} f(t_k) e^{i(2\pi nk)/N} + e^{i(2\pi n(N/2))/N} \sum_{p=0}^{(N/2)-1} f(t_{p+(N/2)}) e^{i(2\pi np)/N}.$$

Voltando o parâmetro p de volta para k e, como n é inteiro, temos que $e^{i(2\pi n(N/2))/N} = e^{i\pi n} = (-1)^n$. Logo,

$$\mathcal{F}(n) = \sum_{k=0}^{(N/2)-1} f(t_k)e^{i(2\pi nk)/N} + (-1)^n \sum_{k=0}^{(N/2)-1} f(t_{k+(N/2)})e^{i(2\pi nk)/N}.$$

Pode-se agora juntar as somatórias e colocar a exponencial em evidência:

$$\mathcal{F}(n) = \sum_{k=0}^{(N/2)-1} [f(t_k) + (-1)^n f(t_{k+(N/2)})]e^{i(2\pi nk)/N}.$$

Considerando os valores pares e ímpares separadamente, tem-se:

$$\mathcal{F}(2m) = \sum_{k=0}^{(N/2)-1} [f(t_k) + f(t_{k+(N/2)})]e^{i(2\pi 2mk)/N}, \qquad m = 0, 1, ..., \frac{N}{2} - 1, \tag{10}$$

e

$$\mathcal{F}(2m+1) = \sum_{k=0}^{(N/2)-1} [f(t_k) - f(t_{k+(N/2)})]e^{i(2\pi(2m+1)k)/N}, \qquad m = 0, 1, ..., \frac{N}{2} - 1. \tag{11}$$

Agora ao invés de uma transformada de Fourier de N pontos, são obtidas 2 transformadas de $N/2$ pontos, e esse processo pode ser repetido transformando em quatro transformadas de $N/4$ pontos, até que se tenha transformadas de Fourier de 2 pontos (Cadzow and Landingham, 1985). Dessa forma, a transformada pode ser calculada recursivamente resultando na transformada rápida de Fourier (em ingles *fast Fourier trasform* ou FFT). Computacionalmente se tem $\log_2 n$ recursões, onde cada uma possui $O(N)$ operações. No total, a FFT realiza $O(N \log_2 N)$ operações, enquanto a tranformada de Fourier de tempo discreto apresenta $O(N^2)$. Na prática essa redução faz uma enorme diferença no tempo de operação para um N grande (Phillips and Parr, 1995).

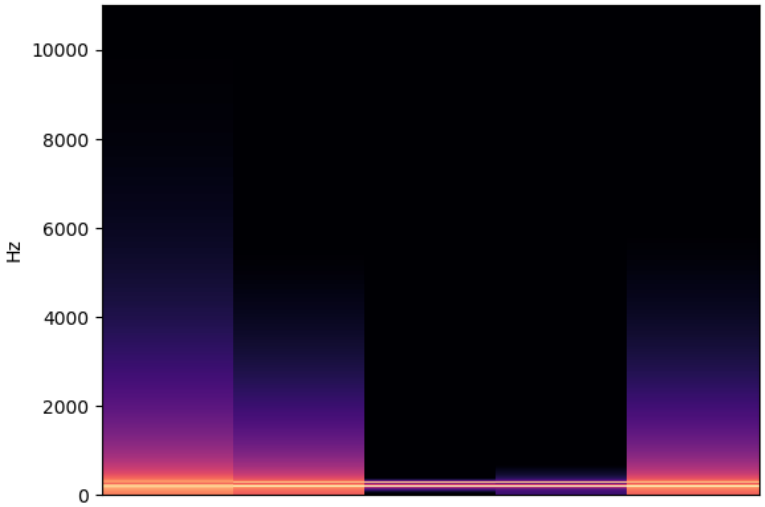

Figura 4: Espectrograma da função $h(t)$

4 Espectrograma

4.1 Definições e Aplicações do Espectrograma

Um espectrograma é uma representação da distribuição de energia de um sinal em termos do tempo e da frequência. O espectrograma é calculado dividindo o sinal em fatias, cada uma delas é caracterizada pela coordenada temporal que se encontra no centro. Multiplica-se cada fatia por uma função de janela (o começo de uma fatia não necessariamente começa ao final de outra) e, após isso, é computada a FFT desse pedaço. O objetivo da função de janela é focar a visão da transformada de Fourier nas proximidades de um determinado ponto, geralmente o central. O resultado do espectrograma é a magnitude ao quadrado da transformada de Fourier. Dessa forma, dado um determinado tempo é possível saber quais são as frequências predominantes. O gráfico do espectrograma possui 3 dimensões. Para representa-lo em 2 dimensões a distribuição de energia (magnitude ao quadrado da transformada de Fourier) é caracterizada por um esquema de cores. (Fulop, 2011). A Figura 4 mostra o espectrograma do sinal $h(t)$ (mostrado na Figura 1). Quando se trata de sinais de som, é normalmente aplicado uma magnitude logarítmica ao espectrograma.

Funções de janela, são funções matemáticas que possuem valor zero fora de um intervalo escolhido (Prabhu, 2014). A função de Hann dada pela equação (12) é um exemplo,

ela vale zero para todos os valores fora do intervalo $0 \leq |n| \leq \frac{N}{2}$.

$$\mathcal{F}(n) = 0,54 + 0,46 \cos\left(\frac{2\pi n}{N}\right), \qquad 0 \leq |n| \leq \frac{N}{2}. \tag{12}$$

A função de janela escolhida é muito importante para as propriedades particulares do espectrograma. É importate que o intervalo entre um tempo central e outro seja menor que o comprimento da função de janela, de preferência menor que o comprimento efetivo (aproximadamente metade do total). É importante também que essas cosiderações sejam seguidas, caso contrário o espectrograma pode não ser uma boa representação do sinal (Fulop, 2011). Por um lado, quanto maior o comprimento da janela melhor a resolução do espectrograma para as frequências, porém sua resolução temporal é pobre. Por outro lado, se o comprimento da janela for pequeno o espectrograma irá possuir uma boa resolução no tempo, mas uma pobre resolução em frequências (Rabiner and Schafer, 2007).

4.2 Motivação e Definição do Espectrograma na Escala Mel

Pitch é uma propriedade do som, que corresponde à percepção humana das frequências. A sensibilidade das pessoas para notar mudanças nas frequências não é linear, por exemplo, humanos possuem maior sensibilidade para certas faixas que para outras. Logo pitch é um atributo subjetivo do som e está relacionado a frequência fundamental (Rabiner and Schafer, 2007). A relação entre pitch e frequência é dada pela escala mel, obtida de forma empírica, pela equação:

$$f_{mels} = 127 \log_e(1 + f_{Hz}/700). \tag{13}$$

Pode-se converter um sinal no domínio das frequências e depois gerar o espectrograma dele. Isso é importante na hora de analisá-lo, pois são dadas importâncias iguais para intervalos que possuem a mesma percepção e não para intervalos iguais entre as frequências.

5 Experimentos Computacionais

Utilizando a nota lá da 4ª oitava gerada em um piano e que possui frequência fundamental de 440Hz, foram realizados testes em `python`. A biblioteca `librosa` possibilitou as

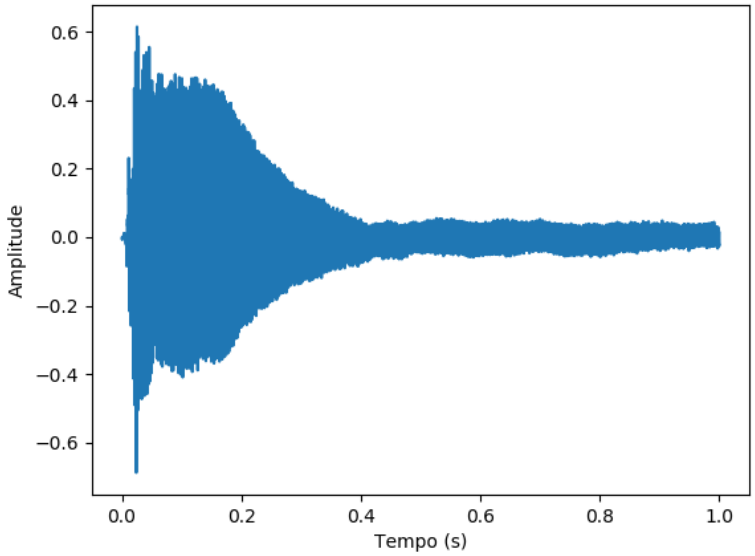

Figura 5: Forma de onda da nota Lá no piano

análises discutidas neste relatório na nota Lá. `Librosa` é um pacote do `python` para análise de músicas e áudios (Librosa, 2020a).

5.1 Lendo Áudio

Dado o sinal de áudio é possível utilizar a função `librosa.load()` para ler o arquivo dentro do `python`. A função `librosa.load()` lê o arquivo de áudio como uma série de dados tipo float. Basta passar o lugar em que o sinal está armazenado no computador e a função então retorna as amostras de áudio e a taxa de amostragem, para mais informações pode-se consultar a documentação (Librosa, 2020b).

Com os dados do áudio é possível gerar o gráfico da forma de onda, utilizando outra biblioteca do python chamada de `matplotlib`. A forma de onda da nota lá pode ser vista na Figura 5.

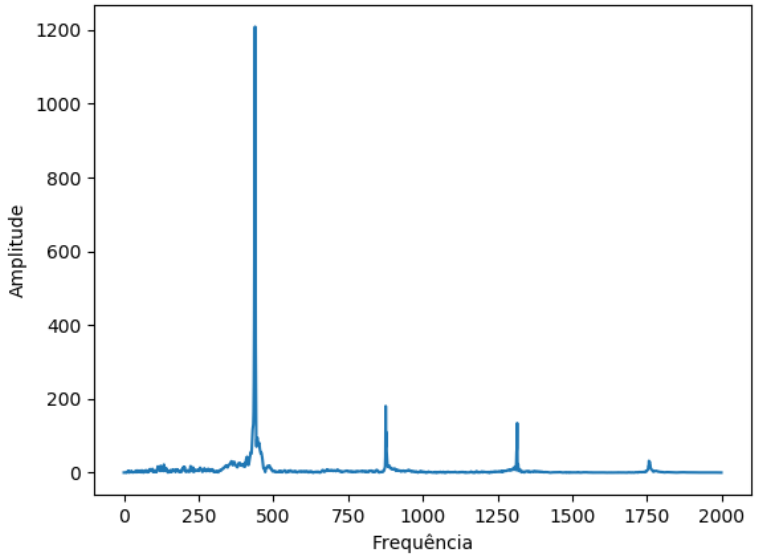

Figura 6: Tranformada de Fourier da nota Lá do piano

5.2 Áudio no Domínio das Frequências

`Numpy` é outra biblioteca do `python` muito utilizada. Com ela pode-se realizar diversas coisas, entre elas a transformada rápida de Fourier do sinal (Numpy, 2020). Para gerar a FFT de um áudio utiliza-se a função `numpy.fft.fft()`, bastando apenas passar as amostras do sinal.

Novamente utilizando a biblioteca `matplotlib` é possível gerar o gráfico da transformada rápida de Fourier do sinal. A Figura 6 mostra a FFT da nota lá, nela pode-se notar um pico na frequência de 440 Hz que é justamente a frequência fundamental da nota, os outros picos correspondem aos múltiplos de 440.

5.3 Gerando o Espectrograma

`Librosa` também fornece maneiras de gerar o espectrograma de um determinado sinal. A seguinte linha de código pode ser usada para gerar o espectrograma do áudio. Para mais informações sobre as funções basta consultar a documentação (Librosa, 2020a).

```
D4 = librosa.amplitude_to_db(np.abs(librosa.stft(amostras_do_sinal),
ref=np.max)
librosa.display.specshow(D4, y_axis='linear', fmax=6000)
```

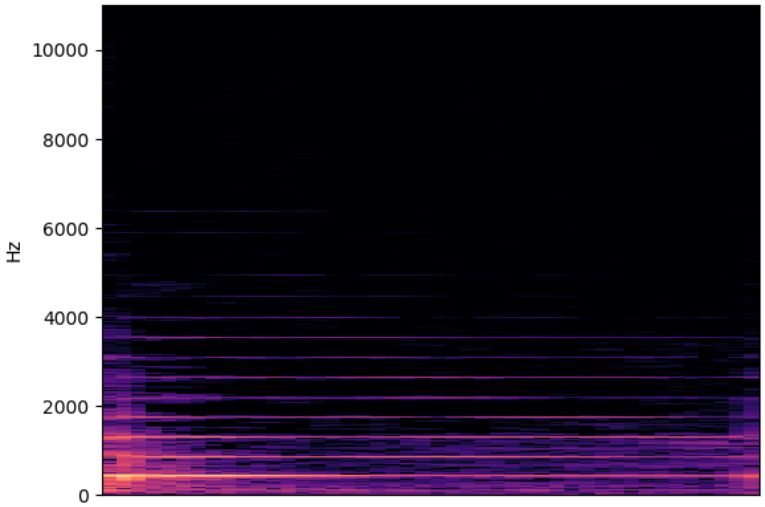

Figura 7: Espectrograma da nota Lá do piano

A Figura 7 mostra o espectrograma da nota lá, é possível notar que existe uma faixa correspondente a frequência de 440Hz indicando que ela está está presente durante todo o sinal.

Novamente utilizando a `Librosa` é possível gerar o espectrograma do áudio na escala mel com o seguinte código:

```
S3 = librosa.feature.melspectrogram(amostras, taxa_de_amostragem)
librosa.display.specshow(librosa.power_to_db(S3, ref=np.max))
```

A Figura 8 ilustra o espectrograma em escala mel da nota Lá do piano.

6 Considerações Finais

Os estudos descritos nessa monografia da disciplina MS777 - Projeto Supervisionado corresponde aos seis primeiros meses de um projeto de iniciação científica com duração de 12 meses. As atividades esperadas para os primeiros seis meses, que inclui o estudo dos Conceitos de Sinais Contínuos e Discretos no Tempo, Transformada de Fourier e Espectrograma, foram devidamente cumpridas e estão descritas na monografia. Nos próximos passos, que correspondem aos próximos seis meses, estudaremos a teoria de Redes Neurais. Sobretudo, aplicaremos os conceitos estudados para rotulação de músicas de uma base de dados apropri-

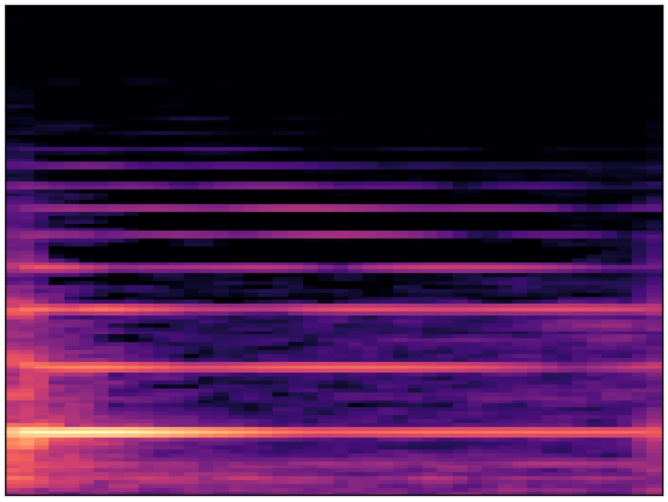

Figura 8: Espectrograma na escala mel da nota Lá do piano

ada. Para isso é necessário primeiramente tratar os dados e só então aplicá-los nas redes neurais para que sejam geradas as rotulações.

Referências

T. Bertin-Mathieux, P.W. Ellis Daniel, B. Whitman, and P. Lamere. The Million Song Dataset. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, 2011.

James A. Cadzow and Hugh F. Van Landingham. *Signals, Systems, and Transforms*. Prentice-Hall, 1985. ISBN 0-13-809542-6.

K. Choi, G. Fazekas, K. Cho, and Sandler M. CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR MUSIC CLASSIFICATION. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA., 2016. IEEE.

K. Choi, G. Fazekas, K. Cho, and Sandler M. The Effects of Noisy Labels on Deep Convolutional Neural Networks for Music Tagging. *Transactions on Emerging Topics in Computational Intelligence*, 2:139–149, 2018.

E. Eck, P. Lamere, T. Bertin-Mathieux, and S. Green. Automatic Generation of Social Tags for Music Recommendation. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*. NIPS, 2008.

Sean A. Fulop. *Speech Spetrum Analysis*. Springer-Verlag Berlin Heidelberg, 2011. ISBN 978-3-642-17477-3.

J. Ker, L. Wang, J. Rao, and T. Lim. Deep Learning Applications in Medical Image Analysis. *IEEE Access*, December 2017.

P. Lamere. Social Tagging and Music Information Rertieval. *Journal of New Music Research*, 2(37):101–114, November 2008.

J. Lee and J. Nam. Multi-Level and Multi-Scale Feature Aggregation Using Pretrained Convolutional Neural Networks for Music Auto-Tagging. *IEEE Signal Processing Letters*, 24(8): 1208–1212, 2017.

Librosa. `https://librosa.github.io/librosa/`, 2020a. Accessed: 2020-02-10.

Librosa. `https://librosa.github.io/librosa/generated/librosa.core.load.html`, 2020b. Accessed: 2020-02-10.

J. Lorince, K. Joseph, and P.M. Todd. Analysis of Music Tagging and Listening Patterns: Do Tags Really Function as Retrieval Aids? In *Agarwal N., Xu K., Osgood N. (eds) Social Computing, Behavioral-Cultural Modeling, and Prediction.*, volume 9021, 2015.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. An Analysis of Neural Language Modeling at Multiple Scales. *arXiv preprint arXiv:1803.08240*, 2018.

Numpy. `https://numpy.org/`, 2020. Accessed: 2020-02-10.

Alan V. Oppenheim and Alan S. Willsky. *Sinais e Sistemas*. Pearson Education do Brasil, 2 edition, 2010. ISBN 978-85-7605-504-4.

Charles L. Phillips and John M. Parr. *Signals, Systems, and Transforms*. Prentice-Hall, 1995. ISBN 0-13-795253-8.

K. M. M. Prabhu. *Window Functions and Their Applications in Signal Processings*. CRC Press, 2014.

Lawrence R. Rabiner and Ronald W. Schafer. Introduction to Digital Speech Processing. *Foundations and Trends in Signal Processing*, 1(1-2):1–194, 2007. doi: 10.1561/2000000001.

J. Shen, M. Wang, S. Yan, H. Pang, and X. Hua. Effective music tagging through advanced statistical modeling. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 635–642, July 2010.

Jayme Vaz Jr. and Edmundo Capelas de Oliveira. *Métodos Matemáticos*, volume 2. Editora da Unicamp, 2016. ISBN 978-85-268-1342-7.

Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan. Facial Landmark Detection with Tweaked Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):3067–3074, December 2018.

S. Zhang, L. Yao, A. Sun, and Y. Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(5:38), 2019.