Três *Datasets* criados a partir de um banco de Canções Populares Brasileiras de Sucesso e Não-Sucesso de 2014 a 2019

André Augusto Bertoni[1], **Rodrigo Pinto Lemos**[2]

[1]Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)
Universidade Federal de Goiás (UFG)
Av. Universitária, n.º 1488 - quadra 86 - bloco A - 3º piso
Setor Leste Universitário - Goiânia - Goiás - CEP: 74605-010

engenheirobertoni@gmail.com, lemos@ufg.br

Abstract. *This work deals with the creation and optimization of a large set of characteristics extracted from a bank of 881 popular Brazilian Hit-Songs and Non Hit-Songs, between January 2014 and May 2019. From this bank of songs, three DataSets were created of distinct features, the first of which contains 3215 statistical features; the second and third are completely new, as they were formed from the Vocal Melody of the songs (Predominant Voice Melody), with no similar database available for research. The second bank represents a spectrogram graph, formed from the initial 90 seconds of each song. The third bank is the most peculiar of all, as it represents a musical semantic analysis of the second bank, where the main purpose was to build a table composed of the most frequent melodic sequences of each song. Our Datasets use only Brazilian songs and focus your data on a limited and contemporary period. The idea behind creating these datasets is to encourage the study of Machine Learning techniques that require musical information. The resources extracted may help in the development of new research in the areas of music and computing in the future.*

Resumo. *Este trabalho trata da criação e otimização de um grande conjunto de características extraídas de um banco de 881 canções populares brasileiras de Sucesso e Não-Sucesso, entre janeiro de 2014 a maio de 2019. A partir desse banco de canções, criou-se três DataSets de características (features) distintas, sendo que o primeiro contém 3215 características estatísticas; o segundo e o terceiro são totalmente inéditos, pois foram formados a partir da Melodia Vocal das canções (Melodia Predominante da Voz), não havendo banco semelhante disponível para pesquisa. O segundo banco representa um gráfico de espectrograma, formado a partir dos 90 segundos iniciais de cada canção. O terceiro banco é o mais peculiar de todos, pois representa uma análise semântica musical do segundo banco, onde a finalidade principal foi construir uma tabela composta pelas sequências melódicas mais frequentes de cada canção. Nossos Datasets usam apenas canções brasileiras e concentram seus dados em um período limitado e contemporâneo. A ideia da criação desses conjuntos de dados é estimular o estudo de técnicas de Aprendizado de Máquina que requeiram informações musicais. Os recursos extraídos podem auxiliar no desenvolvimento de novas pesquisas nas áreas da música e computação no futuro.*

1. Introdução

Há vários anos, temos observado uma profunda mudança na indústria do entretenimento musical. A queda de formatos tradicionais, como os discos físicos (Vinil, CD e DVD) e a ascensão de novos formatos, consumidos exclusivamente através de *streaming*, acabou mudando para sempre os paradigmas sócio-econômico-culturais da forma como ouvimos música em nossas vidas. Plataformas Digitais como Deezer, Spotify e Itunes agora fornecem milhares de músicas à palma de nossas mãos, através dos *smartphones*. Para manter sua base de assinantes o máximo de tempo conectados em suas plataformas, essas empresas vêm aprendendo que devem conhecer a fundo seus produtos (canções) e também seus usuários. É a partir dessa ideia que se começa a dar mais importância aos estudos sobre características (*features*) extraídas diretamente das canções. Essas técnicas, comumente conhecida como MIR (*Music Information Retrieval*), se resumem a extrair-se o máximo de informações estatísticas das canções através do uso de algoritmos especializados. Assim, para cada canção é gerado um conjunto de informações musicais que é armazenado em um banco de dados e posteriormente utilizado pelas empresas de *streaming* em análises estatísticas, verificando, com isso, o desempenho das canções a fim de promover maior engajamento em sua base de usuários [Raieli 2013].

Alguns bancos já foram disponibilizados para a comunidade científica, como é o caso de [Bertin-Mahieux et al. 2011], que promove uma coleção de recursos de áudio e metadados disponível gratuitamente para um milhão de faixas de músicas internacionais populares contemporâneas. Outro exemplo é o [Olteanu 2020], que disponibiliza um banco contendo cerca de 1000 canções, divididas em dez gêneros musicais, apresentando aproximadamente sessenta características de áudio para cada canção. Outro banco bastante utilizado é o [Ay 2018], que propõe a extração de cerca de vinte características de 175.000 canções entre os anos de 1921 a 2020, utilizando-se canções que se destacaram nas melhores posições da revista Billboard Norte Americana. Um detalhe importante a se ressaltar é que, assim como no primeiro banco, estes dois últimos também possuem somente canções internacionais.

Por esse motivo, a inexistência de bancos de dados contendo informações sobre características musicais de canções brasileiras, bem como as enormes diferenças sócio-culturais entre o mercado brasileiro e as demais culturas, foram determinantes para a realização deste trabalho.

Propõe-se, portanto, a criação de três *DataSets* de dados contendo características extraídas de um banco de canções contemporâneas populares brasileiras. As características principais destes bancos é que estão delimitados temporal e culturalmente. Ou seja, a análise foi feita em canções que estiveram em voga no período de Janeiro de 2014 a Maio de 2019. Sendo 441 de Sucesso e 441 de Não-Sucesso.

O primeiro *DataSet* proposto contém 3215 características, tanto de natureza qualitativa quanto quantitativa - com ou sem dependência temporal. O segundo contém a característica "Melodia Predominante", que representa a melodia executada pela voz principal das canções. Já o terceiro representa uma análise semântica musical do conjunto melódico gerado pelo segundo. O algoritmo desenvolvido pela equipe conseguiu verificar e destacar quais são as sequências melódicas mais frequentes em cada canção, utilizando-se o apoio da teoria musical conhecida, levando-se em conta conceitos sobre escalas musicais e campo harmônico de cada canção analisada.

Este trabalho descreve os procedimentos de criação destes três *DataSets*, e está organizado da seguinte maneira: a seção 2 descreve a construção do banco de canções e a formação do primeiro *DataSet* de características; a Seção 3 trata da extração da Melodia Predominante da Voz, pormenorizando a formação do segundo *DataSet*; a Seção 4 explica sobre os detalhes da construção do *DataSet* de informações Semânticas Melódicas, criado a partir do segundo *DataSet*. Por fim, a Seção 4 traz as conclusões do trabalho.

2. Criação do Banco de Canções

A primeira etapa consistiu em realizar um levantamento do conjunto de canções que estiveram nas melhores colocações dentro de um sistema de Ranking e que pudessem ser classificadas como um *Hit* (Sucesso) dentro do período da análise proposta. O principal problema enfrentado nessa construção, foi o fato de que existem diversos métodos que podem ser usados para se mensurar o desempenho das canções, tais como: a) número de acessos por meio de mídias sociais como o Youtube, Instagram ou Facebook; b) *trending topics*[1] do *Twitter;* c) arrecadação e distribuição de direitos autorais pelo **ECAD**[2]; d) número de *Streamings* verificados nas principais plataformas digitais de música: Spotify, Deezer, Itunes; e) Número de execuções dessas canções nas rádios brasileiras dentro do período analisado.

Optou-se por escolher o parâmetro do Número de Execuções em rádios brasileiras, pois este parâmetro ainda proporciona uma maior confiabilidade na aquisição dos dados, já que esse método ainda é a maneira mais comum de se verificar a performance de novas canções por artistas e empresas ligadas ao segmento artístico no mundo todo.

2.1. ConnectMIX

Em toda parte do mundo existem empresas especializadas em monitorar o número de execuções em rádios. No Brasil, esse serviço também é prestado pela empresa **Connectmix** [ConnectMIX 2019], que oferece uma ferramenta homônima de monitoramento em tempo real para auditoria e gestão de *Broadcasting* para emissoras de rádio e televisão. Os dados sobre a performance das canções utilizadas em nosso banco principal foram obtidos através da **Connectmix**, e foram organizados segundo o exemplo mostrado na Tabela 1.

[1]***Trending Topic*** (TT) é tópico em tendência. Isso significa que um grande número de tuítes com uma hashtag ou palavra(s) relacionada(s) a este tópico têm sido disseminada por um grande número de pessoas em um determinado período. Quando isso acontece, o assunto entra para um ranking do *Twitter* de assuntos mais populares e se torna um *Trending Topic* do Twitter

[2]O Escritório Central de Arrecadação e Distribuição (ECAD) é um escritório privado brasileiro responsável pela arrecadação e distribuição dos direitos autorais das músicas aos seus autores, tendo sua sede localizada no Rio de Janeiro.

Tabela 1. Ranking 100+ ConnectMix - Ano 2014

Ano	Posição	Artista	Título	Estilo	Nº Exec.
2014	1º	Marcos e Belutti	Domingo de Manhã	Sert.	384067
2014	2º	Zezé Di Camargo e Luciano	Flores Em Vida	Sert.	273933
2014	3º	Cristiano Araújo	Cê Que Sabe	Sert.	246369
2014	4º	Eduardo Costa	Os 10 M. Do Amor	Sert.	245669
2014	5º	Jorge e Mateus	Calma	Sert.	239337
⋮	⋮	⋮	⋮	⋮	⋮
2014	100º	Fred e Gustavo	Tó Sou Seu	Sert.	59060

Montou-se, dessa forma, um ranking com as 600 canções mais bem posicionadas, considerando-se um intervalo de observação de Janeiro de 2014 até Maio de 2019, com as 100 canções mais executadas de cada ano.

A fim de trazer mais confiabilidade, equilíbrio e robustez, convencionou-se que as canções de **Não-Sucesso** deveriam pertencer ao mesmo conjunto de artistas encontrados na primeira classe (Sucesso); além disso, para que se pudesse assegurar maior confiabilidade na formação do banco das canções de Não-Sucesso, essas canções obrigatoriamente não poderia estar listada em nenhuma posição do Ranking Geral das canções mais executadas, segundo a própria **ConnectMIX**. Para gerar mais equilíbrio entre os dois bancos, optou-se também por selecionar as mesmas quantidades de canções de um mesmo artista para ambos os bancos.

De posse das 600 canções mais tocadas nas rádios nos anos de 2014 a 2019, foi necessário eliminar algumas inconsistências: a) canções que se repetem em mais de um ano de análise; b) versões distintas das mesmas canções (ao vivo ou estúdio); c) músicas gravadas por mais de um artista no período; d) canções internacionais, pois não fazem parte do escopo deste estudo. Dessa forma, chegou-se ao número de 882 canções, sendo 441 rotuladas como Sucesso e 441 como Não-Sucesso.

Portanto, é bom ressaltar que as 441 canções de Não-Sucesso foram escolhidas aleatoriamente, utilizando-se como premissa principal, o simples fato de não listarem como canções executadas nas rádios no período de análise descrito. Sendo que o único cuidado tomado, além do mencionado, foi que os artistas e os números de canções de cada artista verificado no primeiro banco deveriam ser respeitados, para criar mais homogeneidade entre os dois bancos.

2.2. Extração e Tratamento de Características

A fim de se reduzir o esforço computacional de extração das características, convém diminuir o intervalo de tempo de observação de cada canção. Pois, como a melodia, a harmonia e as letras das canções se repetem mais de uma vez ao longo dos registros musicais, poder-se-ia tomar trechos repetidos de análise. Além disso, como esses registros originalmente apresentam durações variáveis, faz-se necessário padronizar a duração do intervalo de tempo de observação para que se produza as mesmas quantidades de características extraídas de cada canção.

Para se definir a duração deste intervalo de tempo, foi realizado um breve estudo estatístico sobre o tempo aproximado das estruturas musicais das canções selecionadas.

Sendo todas as canções concebidas para execução radiofônica, elas normalmente compartilham uma estrutura musical bastante semelhante, composta por: Introdução, Verso A/Semi-Refrão, Refrão A, Solo, Verso B/Semi-Refrão, Refrão Final e Arranjo final, como ilustrado na Figura 1.

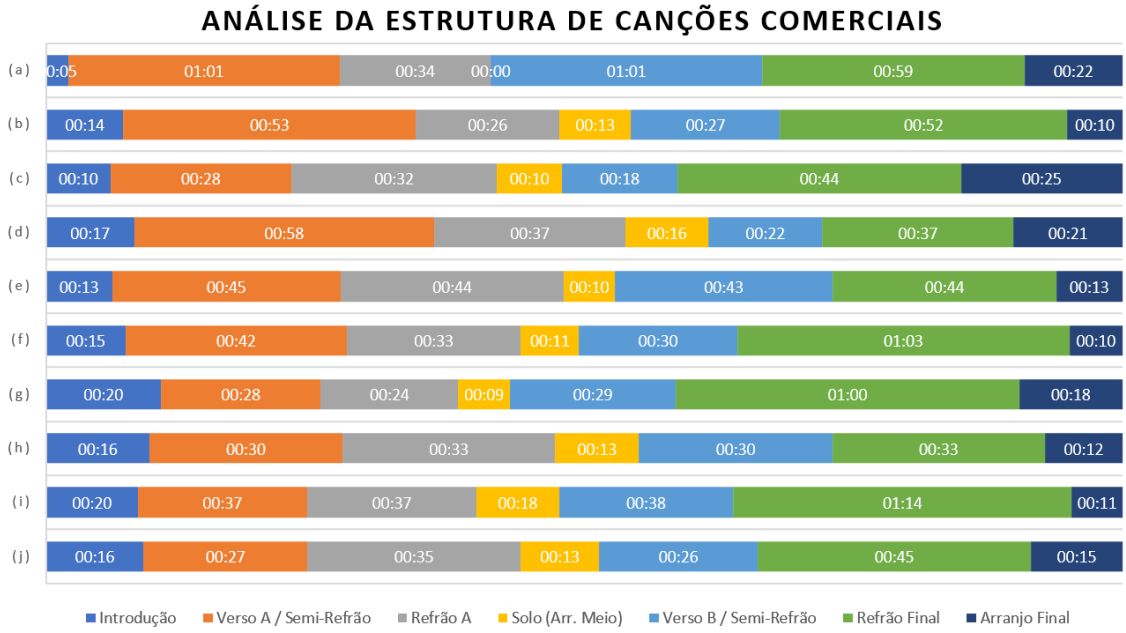

Figura 1. Disposição e duração dos elementos estruturais das Canções de Sucesso (a, b, c, d, e) e de Não-Sucesso (f, g, h, i, j)

Fonte:Próprio Autor

A análise revelou que, em média, os primeiros 90 segundos de cada canção abrangem: Introdução, o Verso A/Semi-Refrão e o Refrão A, que reúnem as principais características de interesse de cada registro musical, pois, após o Refrão A, as canções normalmente repetem os trechos anteriores até o final da música.

Com isso, conclui-se que as músicas só podem ser consideradas tecnicamente inéditas até o final do primeiro Refrão da Canção. A partir deste trecho elas são somente repetições da primeira parte, até que a música chegue ao fim. A Tabela 2 mostrada abaixo, representa o cálculo estatístico das dez canções analisadas na Figura 1.

Tabela 2. Mínimo, Máximo e Média - Estrutura Musical

Duração	Intro	Parte A	Refrão Ini	Arr. Meio	Parte B	Refrão Fim	Arr. Fim
Min	00:05	00:27	00:24	00:00	00:18	00:33	00:10
Max	00:20	01:01	00:44	00:18	01:01	01:14	00:25
Média	00:15	00:41	00:34	00:11	00:32	00:51	00:16

Duração Média das Canções até o fim do 1º Refrão	**01:29**

3. Primeiro *DataSet* de Características de Canções Brasileiras

Para a extração das características das canções, utilizou-se o aplicativo *Streaming Extractor Music*, que compõe o pacote Essentia [Bogdanov et al. 2013]. Essentia é uma biblioteca C++ de código aberto com ligações Python e JavaScript para análise e recuperação de

informações musicais baseadas em áudio. Os aplicativos foram executados em Sistema Operacional Linux (V. 18.04.4 LTS).

Foram desenvolvidos códigos para automatização do processo de extração de características. Como os aplicativos de extração geram como saída um único arquivo ***json*** contendo todas as características desejadas, foi necessário elaborar código que pudesse, carregar, extrair e salvar os arquivos de maneira automatizada.

A Figura 2 mostra uma parte da estrutura do arquivo **json**, apresentando características invariantes no tempo, ou seja, que são calculadas com base em toda a extensão do arquivo de áudio. Os extratores também oferecem características variantes no tempo, calculadas por meio de janelamento, que são apresentadas na Figura 3.

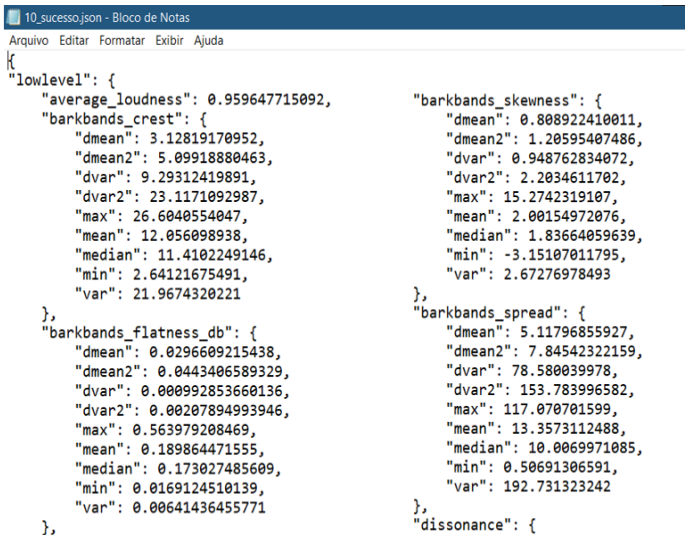

Figura 2. Exemplo de trecho do arquivo json gerado pela ferramenta de Extração do Essentia, no qual figuram algumas *features* e os respectivos valores de seus descritores estatísticos.

Fonte: Próprio Autor

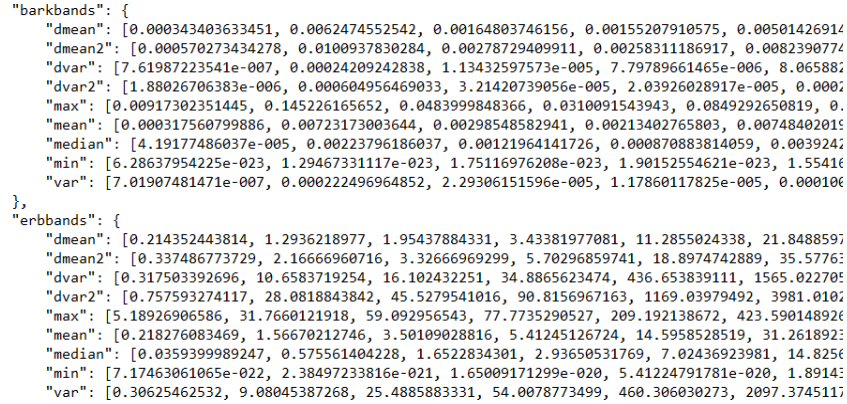

Figura 3. Exemplo de trecho gerado pela ferramenta extratora de característica do Essentia, no qual figuram algumas características e os respectivos valores de seus descritores estatísticos extraídos por Janelamento Temporal

Fonte: Próprio Autor

Por fim, o código extrai características representadas por variáveis categóricas, como mostrado na Figura 4, para o Campo Harmônico da canção ***(chords_key e chords_scale)***, que são descritas por valores do tipo *string*. Como essas variáveis não assumem valores numéricos definidos, utilizou-se a técnica de transformação em variáveis binárias, ou seja, para cada categoria é criado um novo previsor que pode assumir valores binários (0 ou 1). Esse tipo de tratamento é bastante comum em Ciência de Dados, e é conhecido por *Dummy Variables*.

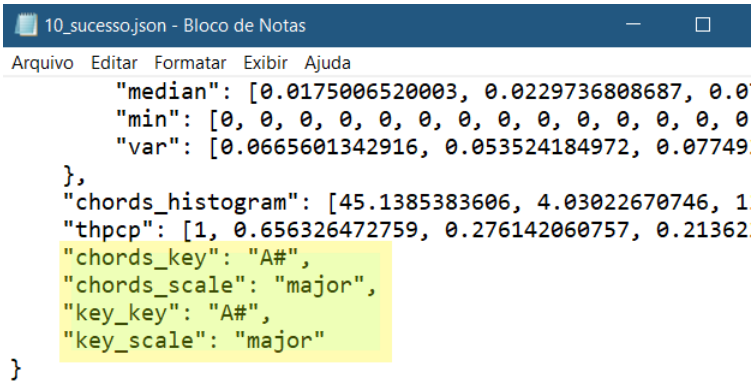

Figura 4. Exemplo de trecho do arquivo json em que foram evidenciadas em amarelo as variáveis categóricas e seus respectivos valores estimados, extraídos através da ferramenta do pacote Essentia

Fonte: Próprio Autor

3.1. *Parsing*

Este processo consiste na extração e reorganização dos dados contidos nos arquivos **json**, convertendo-os em uma tabela, para a qual foi utilizado o formato CSV. O *Parsing* é muito importante nesse processo, pois, somente a partir dessa organização será possível manipular adequadamente os dados obtidos, possibilitando o tratamento das possíveis inconsistências que, em Ciência de Dados, é muito comum e quase sempre necessário.

3.2. O Tratamento utilizando-se a biblioteca Pandas

Após a realização do *Parsing* e a criação do *DataSet* no formato CSV, foi possível carregá-lo no Spyder utilizando-se a biblioteca Pandas.

A princípio foi realizada busca visual, tentando-se encontrar as inconsistências mais comuns. São elas: características indesejadas, dados ausentes (NaN), nulos, divergentes, duplicados, *outliers* e, por último, as variáveis categóricas. Após todos estes tratamentos, pôde-se finalmente formatar a matriz dos dados previsores e o vetor de dados que representa a classe. A Matriz final dos dados ficou da seguinte forma: 882 Linhas (Uma linha para cada Música); 3215 Colunas (Previsores - "Características"); 1 Coluna (Classificador - Sucesso "0"/ Não-Sucesso "1").

O modelo estatisticamente mais eficiente escolhido foi o algoritmo RFE (*Recursive Feature Elimination*), que, após ser utilizado no primeiro banco, conseguiu reduzir as 3215 características para apenas 74 - formando assim o segundo banco.

4. *DataSet* com Melodia Predominante da Voz

Jason Blume afirma que a melodia é a chave principal para o sucesso de qualquer canção [Blume 2019]. Portanto, com base nesta afirmação, é que se propõe a construção de um banco de característica com a informação sobre a Melodia Predominante das canções.

Para extração da melodia predominante, foi empregado o método proposto em [Salamon and Gómez 2012] e implementado e também disponibilizado no pacote Essentia. A melodia predominante é caracterizada pela variação das frequências ao longo do tempo, como apresentada na Figura 5. Dessa maneira, foi gerado um conjunto de dados contendo as linhas melódicas predominantes da voz para cada uma das 882 canções analisadas, as quais também foram armazenadas em formato CSV [Salamon 2013].

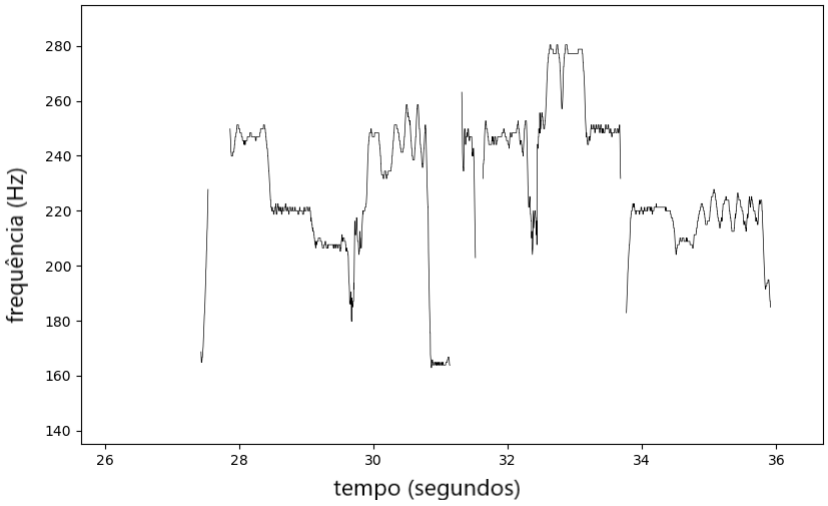

Figura 5. Variação de frequência da melodia ao longo do tempo
Fonte: Próprio Autor

5. *DataSet* com Características Semânticas da Melodia Predominante da Voz

A partir da variação temporal das frequências da melodia predominante, foi desenvolvido um algoritmo que a compara com os campos harmônicos de cada canção para identificar padrões melódicos similares ao longo do intervalo de observação de 90 segundos, de tal forma a extrair informações semânticas das melodias. Entretanto, pode-se observar na Figura 5, que as linhas melódicas apresentam oscilações quase periódicas de frequência, associadas ao efeito de vibrato usado pelos cantores como expressão vocal e apoio para manter a afinação ao longo da interpretação. Além desses artefatos, podem ser identificadas articulações vocais de legato, associadas a ligações entre notas, tanto subindo quanto descendo rapidamente uma escala. Esses efeitos indesejados prejudicam a Análise Semântica Musical devido à grande quantidade de notas fora da melodia principal encontradas em curtos intervalos de tempo.

Para eliminar esses artefatos, primeiramente foi aplicado um filtro passa-faixa para limitar a análise à faixa de 125 a 1500 Hz. que é aquela que concentra maior densidade espectral de potênca e representa melhor as frequências da voz nas canções analisadas. Em seguida, promoveu-se a suavização do traçado da linha melódica da voz calculando-

se sua média móvel ao longo de janelas retangulares de 65 amostras, com deslocamentos de uma amostra. Este procedimento permitiu minimizar as oscilações devidas ao vibrato.

Para remover *outliers* da linha melódica de cada canção, utilizou-se o conceito de escala musical temperada para preservar apenas os valores de frequências abrangidos pelas oitavas completas de C2 (Dó II), 130,812 Hz, até B5 (Si V), 1975,533 Hz. Em seguida, os valores de frequências da linha melódica foram quantizados pelas frequências das notas do campo harmônico de cada canção. Dessa maneira, a linha melódica da Figura 5 assumiu a forma da linha melódica apresentada na Figura 6.

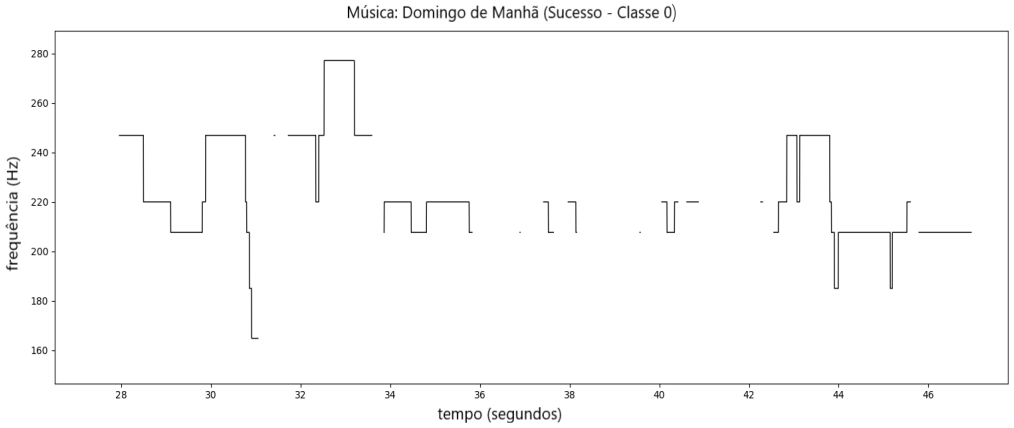

Figura 6. Variação de frequência da melodia ao longo do tempo após a suavização de osilações (vibratos e legatos)

Fonte: Próprio Autor

A partir daqui, foi possível implementar algoritmo que pudesse extrair informações semânticas da Melodia Principal da Voz, realizando uma varredura ponto a ponto em toda a extensão dos 90 segundos da canção. O algoritmo proposto verifica a frequência contida em cada *sample*, criando uma *String* com a sequência de notas musicais encontradas, ignorando as repetições - tomando apenas uma única nota de cada sequência de frequências iguais. O ponto de parada, que delimita o fim de cada sequência melódica é estabelecido quando encontra-se frequências iguais à zero, pois, representam, naturalmente, a pausa de respiração do intérprete naquele trecho da canção. Quando uma nova sequência é encontrada, cria-se, no banco de informações semânticas, nova coluna com o nome da própria sequência melódica, adicionando-se valor 1 à coluna, pois foi a primeira sequência encontrada com aquele padrão de notas. Quando uma sequência já conhecida é encontrada, incrementam-se os valores das colunas.

6. Conclusões

A Ciência de Dados é um segmento da Computação que vem crescendo muito nos últimos anos. A falta da oferta de novos bancos de dados para estudos tem se mostrado sempre um desafio aos pesquisadores, principalmente para os estudos envolvendo Música. O objetivo deste trabalho é colaborar com o desenvolvimento e a criação de novas pesquisas e ou Grupos de Trabalhos que tenham o interesse em auxiliar o desenvolvimento de temas correlatos à música. Certamente, a oferta de novos bancos seria de grande valia para o desenvolvimento de novas pesquisas.

Essa proposta difere bastante de vários outros bancos disponibilizados para estudos na internet, pois, além de focar em um período temporalmente reduzido, reflete uma visão das preferências musicais do maior país da América Latina, que é o Brasil - um país com mais de 200 milhões de pessoas, segundo [IBGE 2021]. Além disso, os outros bancos disponíveis focam principalmente no mercado Norte Americano, contendo centenas de milhares de canções que só abrangem a língua inglesa, ou seja, têm-se disponível poucas informações (características) de um número muito grande de canções, limitado a praticamente um único idioma e mesmo contexto cultural.

Ao contrário de outros bancos de características construídos com outras ferramentas, nossos *DataSets* agregam mais de três mil e duzentas características distintas. O primeiro engloba todas as possíveis características oferecidas pelo pacote Essentia. O segundo oferece características inéditas, que são as melodias predominantes da voz - algo jamais proposto e disponibilizado gratuitamente na internet para estudos. Por fim, o terceiro *DataSet* é o mais inovador, pois, utiliza a sequência melódica da voz para se extrair novas características do ponto de vista musical, e não simplesmente estatístico, como é comum. E isso a partir de código desenvolvido pelos próprios autores, levando-se em conta teoria musical conhecida.

Em trabalhos futuros, pretende-se ainda depurar as informações contidas no banco de melodias vocais, utilizando-se outras abordagens do ponto de vista das teorias musicais conhecidas, tratando-se o problema, por exemplo, sob o ponto de vista de informações conjuntas de harmonia e melodia.

Todos os *DataSets* aqui descritos podem ser encontrados em `https://github.com/tocaestudio/SBBD-2021/`

Referências

Ay, Y. E. (2018). Spotify dataset 1921-2020, 160k+ tracks.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset.

Blume, J. (2019). What makes a song a hit?

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., and Serra, X. (2013). Essentia: an audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR'13)*, pages 493–498, Curitiba, Brazil.

ConnectMIX (2019). Connectmix, monitoramento, auditoria e gestão de áudio em tempo real em rádios e tvs.

IBGE (2021). População do brasil.

Olteanu, A. (2020). Gtzan dataset - music genre classification.

Raieli, R. (2013). *Multimedia Information Retrieval: theory and techniques*. Elsevier.

Salamon, J. (2013). *Melody Extraction from Polyphonic Music Signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain.