UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Tobias Rossi Müller

Um Novo Método para Transcrição Musical Semi-Automática de Áudios Polifônicos com Fontes Múltiplas

Araranguá
2022

Tobias Rossi Müller

Um Novo Método para Transcrição Musical Semi-Automática de Áudios Polifônicos com Fontes Múltiplas

Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Computação do Centro de Ciências, Tecnologias e Saúde do Campus Araranguá da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia de Computação.
Orientador: Prof. Fabrício de Oliveira Ourique, Dr.

Araranguá
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Müller, Tobias Rossi
Um Novo Método para Transcrição Musical Semi-Automática de Áudios Polifônicos com Fontes Múltiplas / Tobias Rossi Müller ; orientador, Fabrício de Oliveira Ourique Ourique, 2022.
22 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Engenharia de Computação, Araranguá, 2022.

Inclui referências.

1. Engenharia de Computação. 2. Transcrição Musical Automática. 3. Recuperação de Informação Musical. 4. Aprendizado de Máquina. 5. Processamento Digital de Sinais. I. Ourique, Fabrício de Oliveira Ourique. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Computação. III. Título.

Tobias Rossi Müller

Um Novo Método para Transcrição Musical Semi-Automática de Áudios Polifônicos com Fontes Múltiplas

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "Bacharel em Engenharia de Computação" e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Computação.

Araranguá, 22 de março de 2022.

Documento assinado digitalmente
Analucia Schiaffino Morales
Data: 25/03/2022 09:11:06-0300
CPF: 622.256.420-87
Verifique as assinaturas em https://v.ufsc.br

Profa. Analucia Schiaffino Morales, Dra.
Coordenador do Curso

Banca Examinadora:

Documento assinado digitalmente
Fabricio de Oliveira Ourique
Data: 25/03/2022 09:39:34-0300
CPF: 916.167.860-00
Verifique as assinaturas em https://v.ufsc.br

Prof. Fabrício de Oliveira Ourique, Dr.
Orientador

Documento assinado digitalmente
Antonio Carlos Sobieranski
Data: 25/03/2022 09:17:02-0300
CPF: 005.305.809-77
Verifique as assinaturas em https://v.ufsc.br

Prof. Antônio Carlos Sobieranski, Dr.
Avaliador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Analucia Schiaffino Morales
Data: 25/03/2022 09:11:33-0300
CPF: 622.256.420-87
Verifique as assinaturas em https://v.ufsc.br

Profa. Analucia Schiaffino Morales, Dra.
Avaliadora
Universidade Federal de Santa Catarina

Um Novo Método para Transcrição Musical Semi-Automática de Áudios Polifônicos com Fontes Múltiplas

A New Method for Semi-Automatic Music Transcription of Polyphonic Audios with Multiple Sources

Tobias Rossi Müller [*] Fabrício de Oliveira Ourique [†]

2022, Março

Resumo

Transcrição musical automática é uma grande tarefa que busca transformar um áudio musical em uma representação simbólica. Essa tarefa é importante porque permite que além de análises no domínio do tempo e frequência sejam realizadas análises semânticas a respeito do conteúdo musical. Neste trabalho, um novo método para transcrição musical polifônica semi automática com instrumentos múltiplos é apresentado, construída a partir de soluções para cada uma das subtarefas que a envolvem. Assim, as diferentes sub tarefas do processo também podem ser avaliadas individualmente, mostrando possíveis gargalos de performance. As técnicas utilizadas para resolver cada sub tarefa envolvem processamento digital de sinais, aprendizado de máquina e conhecimento musical. Os dados utilizados para validar o modelo partiram de um processo de geração sintética a partir de uum banco de amostras. A avaliação de performance foi realizada para o sistema proposto, mostrando forte gargalo no processo de separação das fontes sonoras, e resultado total com uma pontuação F de 0.24, comparável ao 0.28 encontrada na literatura. Algumas críticas são dadas ao processo atual de separação de fontes sonoras, gerando propostas diferentes na abordagem do problema. Por fim, conclui-se que mesmo utilizando diversas soluções para os subproblemas bem estabelecidas na literatura acadêmica e consideradas de boa performance individual, o erro em cascata gerado ainda é muito alto e, caso a separação de fontes sonoras fosse melhor desenvolvida essa performance aumentaria significativamente.

Palavras-chaves: Transcrição Musical Automática. Recuperação de Informação Musical. Separação de Fontes de Áudio. Detecção de Eventos Sonoros. Estimativa de Frequência Fundamental Múltipla.

[*] tobiarossimuller@gmail.com
[†] fabricio.ourique@gmail.com

Um Novo Método para Transcrição Musical Semi-Automática de Áudios Polifônicos com Fontes Múltiplas

A New Method for Semi-Automatic Music Transcription of Polyphonic Audios with Multiple Sources

Tobias Rossi Müller * Fabrício de Oliveira Ourique †

2022, Março

Abstract

Automatic music transcription is a great task that seeks to transform a musical audio into a symbolic representation. This task is important because it allows, in addition to time and frequency domain analysis, semantic analyzes regarding the musical content. In this work, a new method for semi-automatic polyphonic music transcription with multiple instruments is presented, built from solutions for each of the subtasks that involve it. Thus, the different subtasks of the process can also be evaluated individually, showing possible performance bottlenecks. The techniques used to solve each subtask involve digital signal processing, machine learning and musical knowledge. The data used to validate the model came from a synthetic generation process from a sample bank. The performance evaluation was performed for the proposed system, showing a strong bottleneck in the sound source separation process, and a total result with an F score of 0.24, comparable to the 0.28 found in the literature. Some criticisms are given to the current process of separation of sound sources, generating different proposals in approaching the problem. Finally, it is concluded that even using several solutions for the sub-problems well established in the academic literature and considered of good individual performance, the cascade error generated is still very high and, if the separation of sound sources were better developed, this performance would increase significantly.

Key-words: Automatic Music Transcription. Music Information Retrieval. Audio Source Separation. Onset Detection. Multi F0 Tracking.

*tobiarossimuller@gmail.com
†fabricio.ourique@gmail.com

1 Introdução

A Recuperação de Informação Musical (MIR) é uma grande área de persquisa, que luta para resolver diversos problemas que se relacionam a extração de informação a partir de músicas. Ela pode ser dividida em 3 principais pontos focais diferentes, o conteúdo, o contexto e o usuário (KNEES; SCHEDL, 2016). Quando o foco é no conteúdo, as informações são extraídas diretamente do áudio da música. Algumas informações comuns extraídas são o gênero da música, qual o artista ou banda envolvido no trabalho, notas musicais presentes, tonalidade, ritmo, entre muitas outras. Quando o foco é no contexto informações costumeiramente em formato de texto que acompanham o áudio são analisadas. Em foco no usuário uma das principais tarefas e a construção de sistemas de recomendação.

Transcrição musical automática (AMT) é uma tarefa pertencente a MIR baseada em conteúdo, e consiste na transformação de um áudio musical em algum tipo de transcrição simbólica, como cifra e partitura. Uma representação simbólica normalmente é formada pelas notas musicais presentes, com suas respectivas durações e posições, o ritmo da música em batidas por minuto e informações de dinâmica e expressão. O interesse existente nesse tipo de tarefa é devido a que, caso ela seja resolvida, além das análises padrão no domínio do tempo e da frequência de sinais musicais, poderá ser realizada uma análise de nível semântico a respeito do conteúdo. Essas informações semânticas extraídas poderiam ser importantes para indicar o quanto de sucesso que a música pode fazer, quais são as melodias mais curtidas pelo público, ou que se encaixam melhor em dados gêneros musicais, criação de modelos de linguagem musical, entre muitas outras possibilidades. Um tipo de ferramenta existente é a partir da representação simbólica de nível computacional gerar uma de representação de nível humano, como partitura ou cifras, a fim de ajudar músicos amadores ou profissionais a aprenderem a tocar novas músicas mais rápido.

Existem diversas abordagens no desenvolvimento de sistemas de transcrição musical automática, porém uma grande parte delas é voltada a cenários musicais muito específicos, como transcrição sonora de vocais, transcrição especializada em determinado gênero musical, ou até mesmo a transcrição de um áudio de um único instrumento musical, como o piano. Este trabalho propõe um sistema que atua em um cenário mais geral, ou seja, que não depende de gênero, intrumentos tocando ou grau de polifonia. A única informação que o método proposto recebe do usuário de entrada é que a música deve começar logo os primeiros frames do sinal. A polifonia é entendida aqui como a possibilidade de existirem diversas notas musicais tocando simultaneamente, ou seja, diversas frequências fundamentais ao mesmo tempo. Isso acrescenta maior grau de dificuldade, devido a sobreposição de harmônicos e a perda de informação que pode acontecer por efeito de um cancelamento de fase.

O trabalho está organizado da seguinte forma: conceitos e termos importantes para a compreensão do trabalho são explicados brevemente na Seção 2. Em seguida, na Seção 3, alguns trabalhos no mesmo campo de pesquisa são revisados, a fim de compreender com maior detalhes o contexto em que a aplicação está sendo inserida, e quais foram os desenvolvimentos mais importantes e recentes na área de pesquisa. Posteriormente, a metodologia e desenvolvimento do trabalho são relatados Seção 4, juntamente com os experimentos e resultados obtidos. Consequentemente, os resultados dos experimentos são discutidos na Seção 5, tanto com críticas ao sistema como um todo, quanto a seus módulos separadamente. Por fim, as conclusões feitas são demonstradas na Seção 6.

2 Fundamentação Teórica

2.1 Fundamentos Musicais

A seguir conceitos musicais básicos e essenciais para a compreensão e desenvolvimento do projeto são introduzidos.

2.1.1 Notas

O conceito de nota musical é diretamente conectado ao conceito de harmônico fundamental. Instrumentos ditos harmônicos, tem como principal emissão sonora uma onda mecânica composta pela frequência fundamental e seus múltiplos inteiros. Assim, para facilitar a comunicação entre músicos, nomes foram atribuídos as frequências, e se tornaram conhecidas como notas. Naturalmente, instrumentos harmônicos também emitem sons considerados não harmônicos, como barulhos, que também podem ser decompostos em frequências, porém não necessariamente pertecentes aos mútliplos da frequência fundamental.

2.1.2 Escalas

Escalas musicais são conjuntos pré definidos de notas, e estão para a música assim como uma paleta de cores está para a pintura. Nada impede que uma peça musical mude de escala em dado momento, apesar de ser uma técnica pouco comum na música ocidental. Existem diversas famílias de escalas, como maiores e menores. Dentro de uma escala suas notas costumam receber nomes especiais para suas posições, como tônica, dominante e subdominante. Normalmente, escalas recebem o nome da sua nota tônica, seguido pelo nome da família.

2.1.3 Acordes

Assim como escalas musicais, os acordes também são conjuntos de notas, porém, acordes são compostos de notas que ocorrem simultâneamente. Acordes também possuem diversas famílias, que definem o intervalo de semitons entre as notas que o compõe. Famílias de acordes mais famosas são maiores, menores e diminutos (LACERDA, 1967).

2.2 Processamento Digital de Sinais

Diferentes técnicas de processamento digital de sinais como filtros e compressores também podem melhorar o desempenho de um sistema, eliminando informação desnecessária dos sinais de entrada, ou mesmo separando bandas de frequências para diferentes cadeias de processamento.

2.2.1 Autocorrelação

A autocorrelação nada mais é do que a correlação de Pearson sendo utilizada entre um sinal S e sua versão atrasada em N amostras. A autocorrelação se tornou uma ferramenta muito importante para análise de sinais, uma vez que demonstra a influência linear entre amostras anteriores e posteriores.

2.2.2 Transformada Discreta De Fourier em Janelas

A Transformada Discreta de Fourier é uma das ferramenras mais clássicas de processamento digital de sinais, uma vez que processa um sinal de entrada no domínio do tempo e responde com um sinal no domínio da frequência, conjuntamente a fase e a energia contida para cada frequência. A ideia de aplicar a transformada discreta de fourier em janelas de sinal provém da necessidade, porém incapacidade de determinar o local específico do sinal onde uma dada frequência possui mais energia. Essa incapacidade ocorre até mesmo porque o próprio pensamento de procurar por quantidade de energia em uma dada frequência em um dado intervalo vem do mundo contínuo, e deixa de fazer sentido para o caso discreto. Além disso, existe a questão de frequências mais baixas necessitarem de uma maior janela de tempo para poderem ser captadas no processo, seja lembrado aqui a taxa de Nyquist. Assim a ideia de dividir o sinal em janelas menores, garante que ao menos a quantidade de energia calculada para uma dada frequência pertence a janela. Quanto menor a janela escolhida, melhor a precisão de localidade temporal, e pior a capacidade de coletar frequências mais baixas. Além disso, como existirá uma menor quantidade de frames de entrada para a trasnformação, a saída terá o intervalo linear entre as frequências do espectro de saída aumentado. Outro ponto importante na aplicação da técnica é evitar que os sinais sejam divididos em pontos que a amplitude inicial e final sejam maiores que 0. Para evitar tal situação, é comum o uso das chamadas funções de janela, que tem como principal função atuar como um multiplicador de amplitude, que atenuam as janelas de sinal em seu início e fim, levando as amplitudes incial e final do sinal para 0.

2.2.3 Transformada Constante Q em Janelas

A transformada constante Q é uma transformação que quando aplicada a um dado sinal realiza uma transformação do domínio do tempo para o domínio da frequência, assim como a transformação discreta de Fourier. Diferentemente da transformada discreta de Fourier a transforamda Q não tem suas frequências igualmente espaçadas no espectro, pois suas frequências são espalhadas logaritmicamente. Assim, a transformada constante Q pode também ser interpretada como um conjunto filtros passa banda espaçados logaritmicamente. Essa transformação tem sido usada em amplamente para aplicações musicais, uma vez que as frequências fundamentais das notas são também espaçadas da mesma maneira. É observado que apesar da transforamda discreta de Fourier ter mais detalhamento em altas frequências, essa informação é pouco útil, enquanto a trasnformação Q consegue maior detalhamento em frequências mais baixas, informação que tende a ser muito mais útil. A ideia de aplicar a transformada constante Q em janelas vem da mesma ideia de aplicar a trasnformação de Fourier em janelas, apontando quanta energia está distribuída por dada banda de frequência e ganhando uma visão mais temporal de como essas energias se distribuem. Assim como na tranformada discreta de fourier em janelas quanto menor a janela utilizada maior precisão temporal, porém menor precisão das frequências.

2.3 Aprendizado de Máquina

O aprendizado de máquina pode ser compreendido como a capacidade de um computador de encontrar coeficientes para uma dada função, conhecida como modelo, em direção a minimização de uma dada função objetivo, que relacionará a saída do modelo para o conjunto corrente de coeficientes, com os dados reais. Essa função objetivo também é conhecida como função de erro, ou função de perda. Para compreender o grau de generalização do modelo é comum a divisão dos dados em conjuntos de treino e validação,

onde no período de encontrar os coeficientes para o modelo, conhecido melhor como treinamento do modelo, apenas os dados de treino são avaliados pela função objetivo, e em seguida o modelo é avaliado através da função objetivo para os dados de validação. Quanto menor a diferença do cálculo de erro para os conjuntos de treino e teste, melhor é considerada a generalização do modelo. Todo os processo pode ser apenas levado em consideração caso os dados de treino e validação sejam amostrados da mesma população, ou seja, possuam distribuições altamente semelhantes (ABU-MOSTAFA; MAGDON-ISMAIL, 2012).

Existem múltiplas maneiras de buscar por melhores coeficientes para um modelo, e todas elas buscam minimizar uma função objetivo, tornando assim o aprendizado de máquina um processo intrísecamente de otimização. Entre elas, as mais comuns e famosas são, resolução analítica através de igualar a derivada da função a 0, assim encontrando os máximos, mínimos e pontos de sela, o gradiente descendente, que busca sempre ir na direção de mínimos, que não necessariamente são mínimos globais, busca genética ou também pelo algoritmo da colonia de formigas (CHERKASSKY; MULIER, 2007).

Envolta deste processo existem diversos conceitos mais aprofundados, como grau de complexidade das hipóteses do modelo, quantidade de hipóteses existentes para o modelo, capacidade de generalização, métodos supervisionados e não supervisionados, entre outros. Os conceitos citados anteriormente implicam em algumas máximas levadas e consideração e ajudaram a dar direcionamento nas escolhas de soluções durante o desenvolvimento deste trabalho, como viés contra variância. Alguns destes conceitos podem ser analisados pela estatística ou até mesmo pela Teoria da Informação (VAPNIK, 1999).

2.3.1 Aprendizado Profundo

O aprendizado profundo é um subconjunto de modelos baseados em múltiplas camadas de um componente bioinspirado conhecido como neurônio. Neurônios de conectam de camadas anteriores para posteriores, através de uma multiplicação por pesos. Cada neurônio recebe uma dada quantidade de "energia", que será passada para uma função de ativação. Dependendo do nível de "energia"recebido o neurônio pode ou não se ativar. Para o treinamento desse tipo de modelos uma técnica conhecida como retropropagação é utilizada, esta técnica é baseada no gradiente descedente.

Existem diversas arquiteturas para aprendizado profundo. Essas arquiteturas podem ser compreendidas como uma meta estrutura de modelos, ou seja uma questão para o estudo de famílias de funções, topologia matemática. Essas arquiteturas foram criadas pois redes neurais multi camadas comuns, conhecidas como MLP são aproximadores universais de alta complexidade, que implicam em um alto grau de variância. Assim surgiu a ideia de adicionar viéses aos modelos que favorecessem a solução de tarefas estratégicas do dia a dia. As redes neurais convolucionais por exemplo se favorecem do viés de que amostras próximas de um sinal costumam carregar significado semântico em conjunto. Observa-se por exemplo que a imagem de um dado objeto não tem seus pixels com cores completamente aleatórias distribuídos pela matriz, sendo assim, o uso de convoluções ajuda a detectar essa informação conjunta.

2.4 Recuperação de Informação Musical

2.4.1 Transcrição Musical Automática

A tarefa de transcrição musical automática é extremamente complexa, pois depende que diversas subtarefas sejam resolvidas com sucesso para seu pleno funcionamento. Algumas sub tarefas importantes são a detecção de frequência fundamental, que implica na detecção da nota que está tocando em um determinado pedaço de áudio, a detecção de tempo musical e detecção de eventos sonoros, que é utilizada para definir o tempo específico em que uma nota musical ocorre, como também processamentos auxiliares de filtragem, que são importantes para melhorar o desempenho de um sistema como um todo.

2.4.2 Separação de Fontes Sonoras

Um tipo de problema que ganhou visibilidade na comunidade de extração de informação musical é a separação de fontes sonoras. A separação de fontes sonoras tem como objetivo gerar a partir de um sinal S de entrada múltiplos sinais de saída, cada um com uma fonte de áudio diferente contida no sinal S. Diversos sistemas da área focam na extração de vocais de sinais polifônicos, dado o alto valor comercial em aplicações para DJs e produtores musicais. Porém, alguns sistemas procuram resolver a tarefa de forma mais geral, como o Spleeter (HENNEQUIN *et al.*, 2020), aplicação de código aberto desenvolvida utilizando uma arquitetura chamada U-Nets (JANSSON *et al.*, 2017), versão mais específica de redes neurais convolucionais (CNNs). O sistema Spleeter permite a divisão em 4 ou 5 diferentes sinais, sendo as fontes extraídas vocais, percussão, baixo e outros, adicionando extração de piano na versão com 5 tracks.

2.4.3 Detecção de Frequência Fundamental

A detecção de frequência fundamental tem como objetivo compreender que frequência fundamental está tocando em cada trecho específico de um áudio, ou seja, para um dado sinal S de entrada, a saída de um sistema de detecção de frequência fundamental seria uma série F, onde F é amostrada em um dado período de tempo, e tem como valor absoluto a frequência fundamental presente no dado instante de tempo do sinal S original. A definição anterior pode ser estendida para sinais polifônicos, tendo então como saída N séries temporais com os valores das frequências fundamentais detectadas pelo sistema. Essa tarefa também é muito utilizada em um campo de estudo da recuperação de informação musical conhecido como extração melódica, e também na extração de contornos de afinação. A detecção das frequências tem duas principais abordagens, a primeira é o uso das transformadas de Fourier ou da transformada Q, que podem transformar uma janela de áudio específica em um espectro de frequências, enquanto a segunda abordagem também mais recente é pelo uso de modelos de aprendizado de máquina profundo, mais especificamente arquiteturas com camadas Long Short Term Memory (LSTM) e convoluções unidimensionais. Modelos com convoluções unidimensionais com pulos nas conexões são uma possibilidade, baseados na ideia apresentada pelo projeto WaveNet. Um problema clássico ao reoslver a tarefa é de troca de oitavas.

Diversos sistemas ganharam notoriedade na resolução do problema, o primeiro deles a ser citado é o Yin, baseado na técnica de autocorrelação e algumas pequenas correções para resolvê-lo (DE CHEVEIGNÉ; KAWAHARA, 2002). Anos depois o modelo Yin foi aperfeiçoado com a utilização de cadeias de Markov. A ideia da utilização de cadeias de Markov se baseia na premissa de que existe relação entre a frequência de um

instante de tempo e o instante seguinte. Ao ser validado o método gerou resultados que superaram o seu antecessor, e ficou conhecido como o método PYIN (MAUCH; DIXON, 2014). O cálculo do produto interno normalizado entre o espectro do sinal e um cosseno modificado foi uma das soluções importantes para o desenvolvimento da área, e ficou conhecida como SWIPE (CAMACHO; HARRIS, 2008). Com o avanço do aprendizado profundo, devido a crescente capacidade de processamento de dados, o uso de redes neurais convolucionais unidimensionais foi testado para a tarefa, resultando em um ganho de performance significativo quando comparado aos modelos anteriores, este método ficou conhecido como CREPE (KIM *et al.*, 2018).

2.4.4 Detecção de Eventos Sonoros

Detecção de eventos sonoros é uma sub tarefa importantíssima para diversas aplicações de recuperação de informação musical. As utilidades vão de detecção de início de notas musicais a batidas de instrumentos percussivos. O método mais comum de detecção de eventos sonoros é calcular a quantidade de energia presente em uma janela de sinal, e então gerar uma série temporal com as energias das janelas. O tamanho de uma janela é costumeiramente chamado de número de amostras, e o deslocamento de amostras entre cada janela de hop-size. Após a extração da série de energias das janelas um passo de seleção de picos deve ser realizado. Em alguns exemplos a seleção de picos de energia acontece através de um simples limiar de energia, ou de alguma heurística que compare um pico energético e as quantidades de energia próximas ao pico. Exemplos mais avançados costumam utilizar modelos de aprendizado de máquina para a seleção, treinando-os e fazendo seleção de hiperparâmetros através de grandes datasets. Um dos pacotes Python de código aberto que possui solução para essa tarefa é o Librosa (MCFEE *et al.*, 2015), que apresenta uma solução orientada a dados.

2.4.5 Detecção de Tempo Musical

A detecção de tempo musical é uma tarefa útil para a transcrição musical automática devido a importância de definir o ritmo para os músicos tocarem uma dada canção. A mesma costuma ter como entrada uma série temporal S, e como saída um valor único T que corresponde ao ritmo em batidas por minuto detectado pelo sistema. Uma extensão dessa tarefa comum é extrair dinamicamente o tempo musical para cada trecho do áudio, tendo então como saída do sistema uma série T, onde os valores de T representam o ritmo em batidas por minuto em cada instante amostrado. Os instantes amostrados podem ser feitos em um dado período de tempo ou ainda através de uma série de eventos detectada por um detector de eventos. A versão dos sistema é por vezes conhecida como detector de tempo musical dinâmico. A detecção costuma ocorrer através do uso de ferramentas estatísticas como a autocorrelação. A utilização da autocorrelação em uma série de energias de janelas para diferentes quantidades de atraso gera uma distribuição de força de correlação, onde o pico indica a quantidade de amostras de deslocamento. Dada a informação da taxa de amostragem do sinal e de quantidades de amostras de deslocamento no qual a correlação teve seu pico, basta calcular diretamente a quantidade de segundos para a batida, e então converter para batidas por minuto. A solução se aproveita de conceitos musicais para funcionar, como a alta probabilidade de que em cada tempo musical exista algum instrumento percussivo tocando, como a batida de um bumbo, prato ou caixa, e que esses instrumentos costumam conter uma alta quantidade de energia.

3 Trabalhos Correlatos

A busca por trabalhos correlatos aconteceu através de buscas na plataforma IEEE Xplore e também Google Scholar pela palavra chave "Automatic Music Transcription". Uma grande quantidade de trabalhos foi inicialmente selecionada, porém parte deles foi eliminado após a identificação de que não se relacionavam com o tema em questão ou pertenciam a algum cenário da tarefa muito distante do realizado neste trabalho. Por vezes, citações importantes dos artigos selecionados levaram ao conhecimento de novos trabalhos importantes e relacionados com este. Esse tipo de situação ocorreu principalmente durante a leitura dos surveys selecionados ou quando um artigo de proposição de método comparava o próprio com outra solução.

Segundo (GOWRISHANKAR; BHAJANTRI, 2016), existem 3 principais categorias nas quais sistemas de transcrição musical automática podem ser classificados, são elas: transcrição musical informada, transcrição musical de instrumento específico e transcrição musical de gênero específico. O trabalho vai além e faz uma revisão sistemática de trabalhos pertencentes às diferentes categorias apresentadas, demonstrando que dentre eles, as técnicas mais utilizadas são modelos ocultos de markov e máquina de vetores de suporte.

(BENETOS *et al.*, 2018) também traz uma visão geral a respeito de transcrição musical automática, abordando questões como as diferenças entre sistemas baseados em redes neurais e fatoração de matrizes não negativas. O trabalho traz ainda uma visão a respeito de modelos de linguagem musical (MLM), que funcionam analogamente aos modelos de linguagem natural usados na tarefa de reconhecimento de fala. Esses modelos costumam ser utilizados após a etapa de detecção múltipla de frequências fundamentais, que tende a ter um grid probabilístico da existência de cada frequência. O treinamento de um MLM também é análogo ao de um modelo de linguagem natural, sendo que o segundo é costumeiramente treinado a partir de textos de livros, enquanto o primeiro é treinado a partir de um grande conjunto de arquivos midi, que são representações musicais simbólicas. A revisão também trata do entendimento de separação da tarefa de transcrição musical automática em 4 sub tarefas, detecção de frequências fundamentais múltiplas (MPE), transcrição das frequências fundamentais em notas (NT), que depende não só das F0s detectadas, como também da duração e momento em que acontecem. seleção de notas geradas em diversas fontes diferentes de áudio (MPS), e a transformação da representação simbólica em algo humanamente legível, como uma partiurura.

Em (YCART *et al.*, 2019) o uso de um modelo de linguagem musical, baseado na arquitetura LSTM, em conjunto com o uso de um sistema de detecção de múltiplas frequências fundamentais foi capaz de superar o uso de modelos ocultos de markov no conjunto de dados Maps (EMIYA; BADEAU; DAVID, 2009), que consiste em muitas peças musicais clássicas, alinhadas a suas representações simbólicas em formato midi.

O trabalho apresentado por (MCLEOD; STEEDMAN, 2018) explica que não somente a detecção de frequências fundamentais múltiplas é o suficiente para descrever simbolicamente uma música, mas também algumas outras características, como harmônicas e timbre vocal. Segundo os autores, estas informações seriam responsáveis por um processo de transcrição musical mais completo do que apenas as notas musicais aliadas ao ritmo.

(WU; CHEN; SU, 2020) demonstra uma solução para o problema AMT para fontes com múltiplos instrumentos musicais. Além disso, cria 3 definições de cenários importantes para o problema AMT com múltiplos instrumentos, são eles: 1. Instrumento-Informado: cenário onde o sistema recebe como entrada quais são os instrumentos presentes no sinal

de entrada. 2. Instrumento-Agnóstico: cenário onde o sistema desconhece completamente os instrumentos presentes no sinal de entrada. 3. Instrumentos não são transcritos explicitamente na saída do sistema, ou seja, todas as notas musicais detectadas, independente da fonte, vão para a mesma linha de transcrição sem distinção.

O trabalho ainda propõe um método de solução para os cenários Instrumento-Informado e Instrumento-Agnóstico e o compara com outras soluções para cada cenário e dataset específico de avaliação, atingindo resultados próximos ao estado da arte para diversas métricas, como precisão, revocação e escore F.

O artigo escrito por (VACA; GAJJAR; YANG, 2019) cria uma aplicação AMT em tempo real, implementada através de um arranjo de portas programáveis em campo (FPGA). A aplicação apresentada detectava áudio em tempo real através de um microfone e convertia em uma partitura que era apresentada através da tela de um smartphone.

(CHEUK *et al.*, 2021) apresenta uma alternativa de melhoria de para aumentar a acurácia de transcrição através da técnica de reconstrução de espectrograma. O sistema apresentado foi construído totalmente a partir de técnicas de aprendizado não supervisionado, alcançando resultados significativos nos conjuntos de dados MAPS, MAESTRO (HAWTHORNE *et al.*, 2018) e MusicNet (THICKSTUN; HARCHAOUI; KAKADE, 2016).

4 Metodologia

4.1 Modelo Proposto

O método proposto nesse artigo se encaixa em AMT informado com múltiplas fontes, mais especificamente em um cenário Instrumento-Agnóstico. As informações passadas pelo usuário são um ritmo em batidas por minuto próximo a realidade da música passada e a garantia de que as músicas analisadas começam no instante inicial do áudio. A figura 1 mostra um passo a passo estruturado do funcionamento padrão desse tipo de sistema, o qual tirando a ferramenta gráfica, é seguido por este trabalho.

O primeiro procedimento que o sistema realiza é a normalização da amplitude do sinal. Tal processamento é importante devido a facilitar a escolha dos conjuntos de parâmetros dos processamentos seguintes.

Em seguida, o método Spleeter é utilizado em seu modo de 5 canais. A função dele é separar o sinal original em canais de percussão, vocal, piano, baixo em seus próprios canais, os instrumentos não identificados são colocados em um canal específico chamado "outros". A importância desta etapa não está apenas em melhorar a representação simbólica do sistema e sua especificação dos instrumentos, mas também em tornar os diferentes sinais de saída menos polifônicos. A diminuição da polifonia contribui para um melhor desempenho dos processamentos posteriores. A figura 2 demonstra como funciona a tarefa para um exemplo hipotético de 4 canais, sendo eles piano, vocal, percussão e guitarra.

A partir do momento em que os canais estão divididos, dois filtros são aplicados sobre o sinal, a fim de compensar um efeito descrito pela ISO 226:2003 (ISO. . . , 2003), que demonstra uma curva do quanto o ser humano percebe o som, sendo o eixo horizontal a frequência da emissão sonora e o eixo vertical a amplitude necessária para que a percepção sonora seja estável. A curva pode ser observada na figura 3. O primeiro filtro utilizado foi um passa alta de segunda ordem, com frequência de corte em 150 Hertz. O segundo filtro utilizado foi um passa baixa de décima ordem, a uma frequência de corte de 10000 Hertz. As respostas em frequências dos filtros podem ser vistas nas figuras 4 e 5.

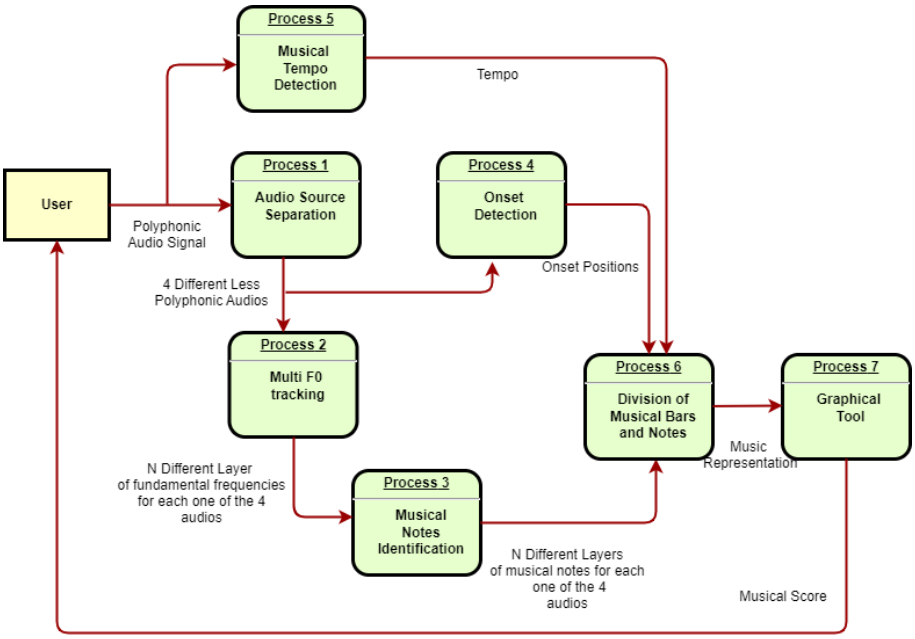

Figura 1 – Demonstração geral das etapas de processamento do sistema apresentado.

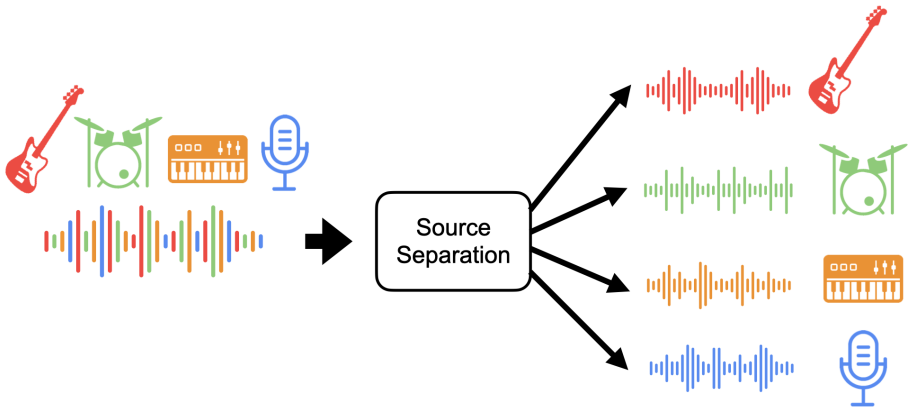

Figura 2 – Demonstração hipotética do processo de separação de fontes sonoras. Disponível em: <https://source-separation.github.io/tutorial/intro/src_sep_101.html> Acesso em: 20 de março de 2022.

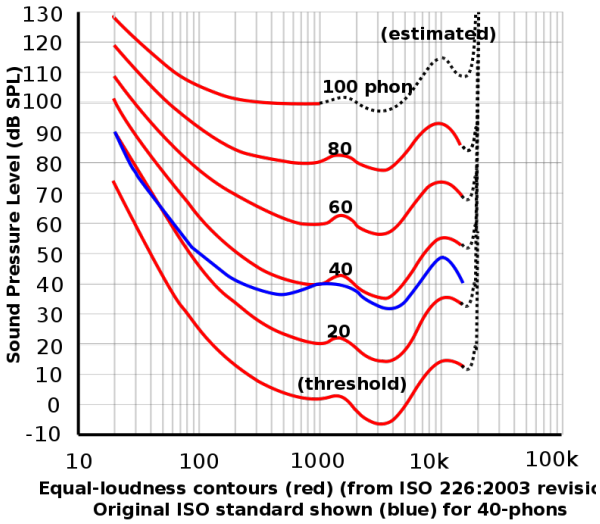

Figura 3 – ISO 226:2003.

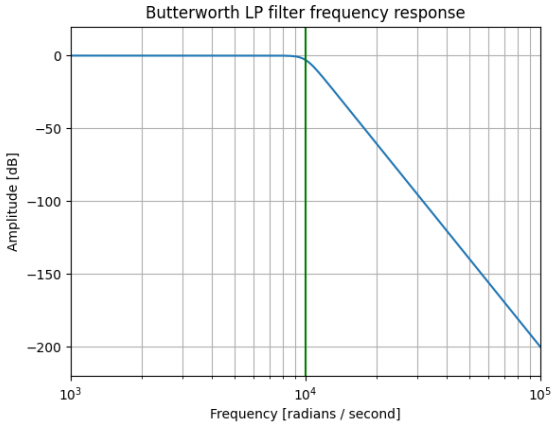

Figura 4 – Filtro passa baixa de décima ordem com frequência de corte em 10000 Hertz.

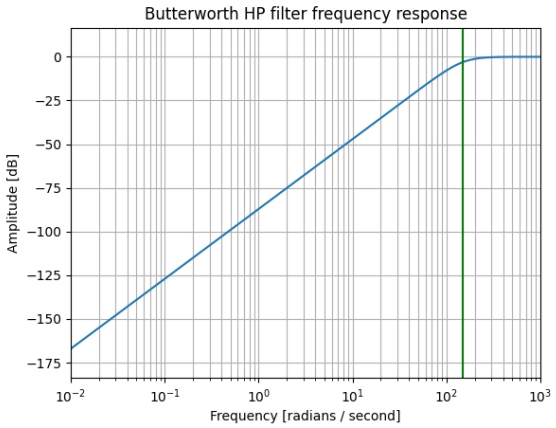

Figura 5 – Filtro passa alta de segunda ordem com frequência de corte em 150 Hertz.

O processo de detecção de frequências fundamentais múltiplas é então iniciado. Observa-se que sistemas desse tipo podem ser divididos em 3 classes principais: extração iterativa, extração linear e extração não linear. Na extração iterativa algum sistema de detecção monofônica seria utilizado e então os múltiplos da frequência fundamental detectada seriam eliminados com filtros de entalhe. O processo se repete iterativamente até que uma dada condição de parada seja alcançada. Na extração linear algum método linear é utilizado de base para decompor o sinal, como em alguns trabalhos que utilizam fatoração por matrizes não negativas. Em extração não linear normalmente o sistema é construído com um conjunto de filtros, e então algum modelo que utiliza o sinal processado como carcterística para prever as frequências fundamentais presentes. Esse tipo de sistema aliado a redes neurais convolucionais caracteriza o estado da arte, chegando a 0.737 de F-score no conhecido dataset MusicNet (WU; CHEN; SU, 2019).

O sistema proposto se encaixa na categoria de extração não linear. Inicialmente a transformada constante Q foi aplicada a janelas de 512 frames. A frequência míinima da transformada foi de 65.41 Hertz, isto é, a frequência fundamental da nota dó na oitava de número 2. A implementação utilizada foi a da biblioteca Librosa, que tem como parâmetro também a quantidade de filtros por oitava. O valor escolhido foi 12, pois cada oitava musical tem 12 semi tons. Ao todo 84 filtros foram utilizados, gerando uma capacidade de pegar harmônicos de 7 oitavas diferentes, isto é, entre a segunda e a nona oitava. Esse processo gera as cartacterísticas iniciais para a resolução da sub tarefa. Em sequência, as características são normalizadas entre 0 e 1.

Considerando que o espectro comum de frequência de uma nota musical é constituído por sua frequência fundamental e múltiplos inteiros positivos, a ideia utilizada foi criar uma nova característica que leve essa informação em conta. Para algumas diferentes carcterísticas foram calculadas e brevemente comparadas. A característica que obteve um maior grau de sucesso nas comparações foi a soma da energia captada no filtro k somado ao k + 12, isto é, a fundamental somada a sua oitava multiplicado a um coeficiente C. O coeficiente C é calculado através da multiplicação das energias captadas nos filtros K + n, onde n pertence ao conjunto de elementos 19, 24, 28, 31. A explicação lógica por trás dessa caracterísitica calculada é que, na grande maioria dos casos, se uma dada nota está tocando, uma boa parte da energia dela está distribuída entre as frequências fundamental e sua oitava. Quanto a multiplicação pelo coeficiente, a explicação é que em certas situações expecíficas pode existir duas frequências f e 2f, e caso os harmônicos seguintes 3f até 6f não existam, talvez seja porque f não seja fundamental.

Uma vez que esse novo vetor de características foi calculado, um algoritmo de seleção de picos máximos pode ser utilizado para selecionar as posições com maior valor e que consequentemente tem maior probabilidade de serem verdadeiros positivos na detecção das notas. Apesar de todos os picos máximos serem selecionados pelo algoritmo apenas aqueles que ultrapassarem um dado threshold são selecionados como as frequências fundamentais presentes na janela de áudio. Todo esse processo é realizado para todas as janelas individualmente, gerando assim como saída um vetor de comprimento igual a quantidade de janelas, que guarda em cada posicão uma lista das frequências fundamentais detectadas na dada janela.

A parte de identificação de notas faz a conversão da frequência fundamental detectada para a nota que a representa. A tabela de frequências fundamentais por nota utilizada tinha como afinação o lá da escala número 4 em 440 Hertz. A implememtação utilizada foi de uma tabela hash que tem como entrada a posição do vetor e como saída

uma string com o nome da nota com sua respectiva oitava.

Com as frequências fundamentais para cada janelas detectadas, o próximo passo foi a detecção de eventos sonoros, que podem ser desde o início ou o fim de uma nota até a reprodução de algum evento não harmônico como por exemplo a batida de um tambor ou prato. Para essa tarefa foi utilizado um algoritmo pré pronto, conhecido na literatura e implementado pela biblioteca Librosa, baseado em alteração espectral entre janelas de aúdio. A ideia consiste em dividir o sinal em janelas, realizar uma transformação do domíni do tempo para o domínio da frequência, e então calcular a diferença entre os espectros da janela atual para a janela anterior, criando um vetor de característica com o tamanho igual a quantidade de janelas menos 1. O espectro utilizado para análise da diferença é com potência em escala logarítmica e frequência em escala MEL, que também é logarítmica. Neste projeto o espectro todo foi analisado para detecção de eventos, porém não é incomum o uso da divisão do espectro em bandas, detectando assim, eventos em cada diferente banda de frequência. Apesar do algoritmo proposto não dividir em banda de frequências, ele é aplicado para cada canal dividido pelo algoritmo de divisão de fontes sonoras, assim conseguindo detectar a reprodução de notas de cada instrumento separadamente. Com o vetor de diferenças espectrais calulado, um algoritmo de seleção de picos máximos é executado sobre o vetor, detectando assim, os pontos exatos em que o evento acontece. A figura 6 demonstra o funcionamento do processo de detecção de eventos sonoros implementado pela biblioteca librosa.

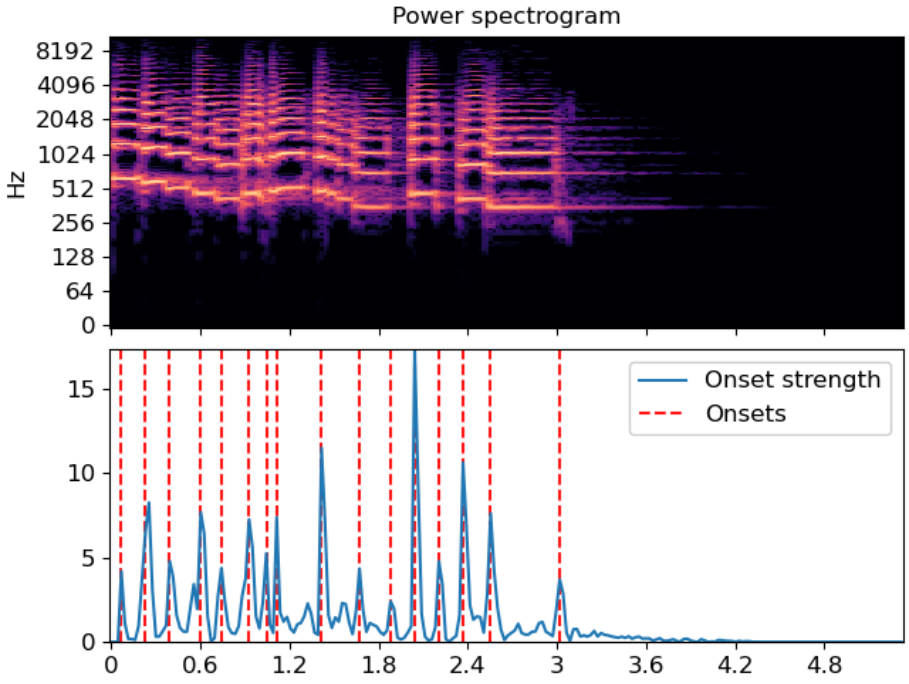

Figura 6 – Demonstração da extração de eventos sonoros de uma amostra de aúdio pela implementação da bilbioteca Librosa. Disponível em: <https://librosa.org/doc/latest/generated/librosa.onset.onset_detect.html> Acesso em: 10 de março de 2022.

A detecção do ritmo ou tempo da música é realizada utilizando a autocorrelação desse vetor de diferenças para diferentes atrasos. A partir das correlações realizadas para diferentes atrasos um vetor de características é calculado. Novamente um algoritmo de

detecção de picos máximos é utilizado, porém para que a detecção fique próxima ao esperado por que está executando a função, e não em algum múltiplo inteiro positivo desse valor, a curva passa por uma modificação baseada em uma distribuição de probabilidade pré definida. A distribuição utilizada nesse projeto foi a log normal com centro em 80 batidas por minuto. Além disso, caso o ritmo detectado fosse menor que um dado valor muito menor do que o centro da distribuição o retorno é calculado multiplicando-se o valor calculado por 2. O sinal do qual o bpm foi extraído foi o não dividido por fontes e normalizado. O motivo disso é que o ritmo é uma cartacterística que todos os intrumentos de uma música seguem em conjunto e, portanto, todo pico de todo instrumento é importante para sua detecção.

Um processo intermediário de extração de amplitude para cada faixa de áudio separada foi realizado, para ajudar no processo posterior de detecção das notas e suas durações. Com as informações de tempo, janelas, amplitude e frequências detectadas para cada janela de cada faixa, o processo de detecção de notas é iniciado.

Se a princípio uma nota é um evento, então basta olhar para as janelas que acontecem eventos para detectar o início das notas. O final da nota é detectado através do envelope de amplitude, isto é, percorre-se da janela do evento no envelope de amplitude até o ponto em que, ou acontece um novo evento, ou a amplitude é baixa o suficiente para ser considerada irrelevante e a nota tenha terminado. Assim são definidas as durações em janelas, e as janelas início e fim de uma nota. Para saber qual nota específica em questão de frequência foi detectada, o intervalo de janelas é analisado, verificando quais são frequências fundamentais predominantemente detectadas. Para que uma frequência fundamental seja determinada como existente nesse período de janelas de áudio, ela deve aparece em pelo menos 30% das janelas do intervalo. Assim a detecção de notas com suas frequências, janela de início e duração em janelas é concluída.

Para converter de janelas para tempo são utilizados o ritmo de amostragem, tamanho de janelas e ritmo da música. para que o tempo detectado não seja um valor quebrado existem aproximações para o quarto de tempo mais próximo.

4.2 Geração de Dados Sintéticos

A ideia de dados sintéticos não é nova, e já é amplamente utilizada na pesquisa em visão computacional, principalemnte através de simulações com motores gráficos famosos, conehcidos como "Engines". Normalmente esse tipo de abordagem é utilizada quando há dificuldade de se adiquirir dados reais para o problema proposto, ou os dados reais são de baixa qualidade. No mundo do áudio existe um grande problema devido a faixas de áudio produzidas por grandes estúdios serem registradas e não poderem ser utilizadas em pesquisa acadêmica, por isso, é muito comum o uso de músicas caseiras, indie ou de estúdios menores para a criação de conjuntos de dados. Além disso, mais especificamente no ramo da trasncriação musical automática, existe a dificuldade associada ao alto trabalho de colocar músicos profissionais fazerem essas transcrições, que muitas vezes são até mesmo vendidas por altos valores na internet. Além disso, trasncrições realizadas por músicos poderiam conter erro humano associado, dificultando o processo de modelagem dos dados.

Considerando que os dataset apresentados na literatura até então não se adequavam perfeitamente a tarefa apresentada pelo método proposto, a solução encontrada foi a criação de um algoritmo de geração de áudios que pudesse criar automaticamente a partir de alguns parâmetros e uma semente randômica um áudio com intrumentos de todas as categorias separadas pelo algoritmo de separação de fontes, além da representação simbólica direta.

O algoritmo criado também tem como saída o áudio gerado separado pelas fontes sonoras em outros áudios.

Para construir esse tipo de sistema, dois caminhos principais poderiam ter sido seguidos. O primeiro deles seria sintetizar puramente todos os intrumentos que seriam utilizados. O segundo seria criar um banco de amostras de cada instrumento, e então criar um áudio completo a partir desse banco de amostras. A segunda opção é mais viável e possível de alcançar bons resultados, pois evita entrar em questões complexas como sintetização de vocais.

O funcionamento da amostragem se baseia no conceito de que notas de instrumentos musicais tem um início denominado ataque bem definido a nível de timbre e em sequência estabilizam em um formato de onda específico e harmônica até que sua amplitude chegue a zero. Assim as notas podem ser posicionadas aleatoriamente sobre um áudio vazio, com duração também aleatória, sendo que caso o áudio exceda o tamanho da duração da nota a amplitude da nota pode ser diminuída levemente até o final da duração da nota aleatorizada.

O sistema de geração do conjunto de dados apresentado considera um super conjunto do que poderia ser considerado música, uma vez que as notas aleatorizadas não foram condicionadas a escalas musicais específicas, ou ainda não levam em conta conceitos importantes de composição como apinhamento harmônico. Outros pontos importantes é que a aleatorização não leva em conta conceitos de harmonização e harmonia funcional, ou seja, são apenas notas aleatórias sobre um áudio vazio, condicionadas a um ritmo específico também aleatorizado.

Os intrumentos utilizados para o preenchimento de todas as classes do separador de fontes foram Piano, Guitarra Elétrica, Baixo de Corda, Vocal e Percussão. Para que o processo de geração da percussão fosse mais condizente com a realidade uma característica extra foi adicionado a geração dos áudios. No meio de produção musical uma técnica muito comum para composição de sequencias percussivas é o uso de sequências pré prontas, normalmente usadas de base para composição de algum gênero musical específico. Então para que a percussão fosse menos aleatória, uma sequência percussiva base foi construída adaptando-se a 3 ritmos diferentes de 80, 100 e 120 batidas por minuto.

Assim um banco de amostras sonoras de cada instrumento utilizado foi criado. As amostras seguem um padrão inflexível de serem apenas áudios do instrumento tocando uma única nota por um tempo de aproximadamente dois segundos. A nota deve começar exatamente no início da amostra. Cada amostra deve ser coloca em uma pasta com o nome específico do instrumento que está sendo tocado, e além disso ser nomeada com o nome da nota tocada e sua respectiva oitava.

Para gerar os datasets para os experimentos com o sistema porposto, 400 áudios diferentes foram criados, com suas respectivas separações por fonte e representação simbólica. Os áudios são de 4 compassos de 4 tempos e possuem 16 notas aleatorizadas, além do sequenciamento de bateria pré programado.

4.3 Experimentos e Resultados

Foram realizados dois experimentos principais com o sistema. O primeiro deles foi avaliar o sistema construído como um todo, ou seja, tendo como entrada o áudio contruído com os diversos intrumentos e analisando a representação simbólica de saída. O segundo experimento realizado foi com os dados separados por fonte perfeitamente pelo gerador, e

Experimento 1			
Instrumento	Precision	Recall	F Score
Piano	0.38	0.17	0.24
Vocal	0.31	0.46	0.37
Baixo	0.54	0.13	0.22
Outros	0.08	0.39	0.13

Tabela 1 – Métricas por instrumento para o experimento 1.

Experimento 2			
Instrumento	Precision	Recall	F Score
Piano	0.63	0.95	0.76
Vocal	0.43	0.82	0.56
Baixo	0.38	0.97	0.54
Outros	0.23	0.70	0.35

Tabela 2 – Métricas por instrumento para o experimento 2.

tem como objetivo compreender o quanto da performance pode ser perdida na etapa de divisão de fontes. Os resultados de detecção de notas por instrumentos podem ser visto nas tabelas 1 e 2. As métricas utilizadas são as padrão da literatura de aprendizado de máquina em problemas de classificação. A métrica "precision"é definida como a quantidade de verdadeiros positivos divididos pela soma entre verdadeiros positivos e falsos positivos. Quanto maior o valor, melhor a precisão do sistema, e quanto menor, naturalmente quer dizer que existe uma quantidade muito alta de falsas detecções de notas. Essa métrica costuma ser bastante prejudicada pelo problema das oitavas, que é quando uma nota é detectada na oitava errada, ou ainda a nota correta e sua oitava superior são detectadas conjuntamente. A métrica "recall"é definida como a quantidade de verdadeiros positivos dividida pela soma entre os verdadeiros positivos e os falso negativos. Sendo assim, quanto maior seu valor, menos detecções que deveriam ter sido feitas são esquecidas pelo sistema. Como a precisão é debilitada por detecções demais, e a revocação é debilitada por detecções de menos, uma métrica foi criada para balancear esses dois tipos de erros, sendo chamado de "F score". A pontuação F é definida como 2 dividido pela soma dos inversos de precisão e revocação. Naturalmente, quanto maior essa métrica melhor o resultado do sistema em balancear os dois tipos de erros.

Para uma detecção de nota ser considerada um verdadeiro positivo ela deve ter duas características entre a representação base e a predição, começar com uma diferença de instante menor que 5 janelas e ter a mesma frequência fundamental.

A duração da nota é avaliada individualmente apenas para os verdadeiros positivos, em unidade de erro médio de tempos musicais. Além disso, também são calculados os erros de início da nota em unidade de tempos musicais e o erro médio do detector de ritmo em batidas por minuto. Essas informações podem ser analisadas na tabela 3.

Experimentos	Mean Absolute Start Error	Mean Absolute Duration Error	Mean Absolute BPM error	Median Absolute BPM Error
1	0.26	0.38	3.55	0.61
2	0.34	0.18	4.10	0.62

Tabela 3 – Métricas gerais de erro para os experimentos 1 e 2.

5 Discussão

Apesar da tarefa de transcrição musical automática já ser explorada na literatura acadêmica, o cenário instrumento agnóstico e semi automático apresentado neste trabalho ainda não foi, sendo assim possível de comparar apenas com cenários relativamente semelhantes, ou ainda de comparar seus módulos individuais com outras soluções existentes. Assim como no trabalho aqui proposto, (WU; CHEN; SU, 2020) também realizou uma proposta de sistema de transcrição musical automática no cenário instrumento agnóstico, porém testando em 3 diferentes conjuntos de dados. O F score médio obtido pelo trabalho foi de 0,46 com o melhor modelo na tarefa análise de frequências fundamentais nas janelas, tarefa específica não avaliada no trabalho proposto, e F score de 0,28 para detecção a nível de notas, contra 0,24 obtido no trabalho proposto no experimento 1 e 0,55 no experimetno 2. Apesar dos resultados resultados obtidos serem muito aprecidos a nível de métricas, é difícil ter certeza como os sistemas performariam um nos dados testados pelo outro. Outros trabalhos são revisados com métricas de performance semelhantes no survey de (GOWRISHANKAR; BHAJANTRI, 2016), porém as tarefas são menos parecidas e mais específicas do que a realizada neste trabalho.

Os resultados dos experimentos mostram problemas sérios com o algoritmo Spleeter em dividir as fontes sonoras, comprometendo drasticamente a performance da aplicação. Aparentemente, o modelo Spleeter tem dificuldades em compreender também que um baixo de corda deveria ser colocado no canal baixo. As detecções corretas analisadas do modelo para essa classe foram quando o baixo era substituído por um grave de sub. Além disso, o timbre de piano utilizado tinha parte de sua potência de sinal colocada para o canal de outros instrumentos, indicando que o algoritmo pode ser bastante sensível a timbres um pouco diferentes do esperado.

O algoiritmo de Spleeter é baseado em decomposição do sinal original com filtros de convolução unidimensionais, e analisa as janelas de forma única, sem compreender o contexto do sinal de entrada. Como crítica ao algoritmo, pode ser que talvez não seja a forma mais eficiente de tratar o problema, uma vez que uma técnica muito comum de composição musical é juntar timbres diferentes tocando a mesma nota no mesmo instante de tempo, para criar um som mais rico e com camadas, que forma um novo timbre, criando uma dependência importante de contexto.

Talvez uma forma mais interessante de tratar esse tipo de problema seja fazendo uma análise do sinal como um todo, afim de tentar detectar pontos de cancelamento de fase entre diferentes janelas do sinal. Um exemplo simples seria caso uma nota de piano e de um trompete sejam tocadas em dados momentos diferentes de um sinal, e em outro momento sejam tocadas conjuntamente. A informação anterior dos timbres separados seria relevante para a separação no momento em que os timbres estivessem tocando em conjunto e formando um novo timbre. Talvez essa forma de tratar o problema o torne

mais robusto para diferentes timbragens, apesar da necessidade posterior de classificar os timbres individualmente identificados.

Existem críticas a serem feitas a todos os módulos apresentados do sistema proposto, e também ao sistema como um todo. Começando pelo sistema como um todo, pode ser observado que em muitos momentos do desenvolvimento limiares foram escolhidos para a seleção de picos máximos, de forma completamente arbitrária e baseada em análise exploratória dos modelos em amostras separadas. Isso pode indicar que os hiper parâmetros escolhidos para o modelo não são os melhores, debilitando sua performance. Para resolver tal problema, poderia ser performada uma etapa de otimização não linear bayesiana. Esse processo seria de altíssimo custo computacional, pois cada inferência de 200 áudios leva em torno de 2 horas para ser executada em processo único, e por isso não foi realizado neste trabalho.

Quanto as partes separadas, observa-se que o processo de modelagem do filtro para equalização do loudness poderia ter sido muito mais complexa, tornando o processo de detecção de múltiplas frequências fundamentais posterior muito mais eficiente.

Quanto ao processo de detecção de múltiplas frequências fundamentais, outra abordagem havia sido inicialmente proposta a partir do método CREPE, com algumas pequenas modificações no cálculo da última camada do sistema. Porém, como o modelo foi treinado para áudios monofônicos, ao ser testado com polifonia teve queda de precisão e performance muito significativa. O módulo foi então trocado por uma extração de características com a tranformada constante Q, e com posterior heurística desenvolvida. A heurística desenvolvida poderia ter sido melhor explorada, levando a possível ganho de performance, e menos problemas com a seleção da oitava. Além disso, como demonstrado nos trabalhos correlatos, essa heurística poderia ter sido substituída por algum modelo preditor de aprendizado de máquina. Porém, o experimento 2 mostrou um score F muito elevado para as notas do piano, sendo comparável até mesmo ao estado na arte alcançado no dataset MusicNet. Vale ressaltar que o alto valor de revocação com valor inferior de precisão pode indicar excesso de detecções de notas, e o aumento do limiar de inclusão da nota poderia trazer ganhos tanto para a métrica de precisão quanto a pontuação F.

Observa-se também que a tarefa de detecção de múltiplas frequências fundamentais pode ser vista como análoga a tarefa de detecção de fala, e processamento de linguagem natural. Em processamento de linguagem natural, é muito comum o uso dos chamados modelos de linguagem. Os modelos de linguagem são pós processadores da predição, que levam em conta questões como a probabilidade de uma palavra vir após a outra. O uso de cadeias de Markov é padrão para a realização dessa tarefa, assim como redes neurais com recorrência. Como na música certas notas tendem a vir depois de outras, existe a possibilidade de explorar modelos de linguagem musical, a fim de melhorar a performance desse tipo de sistema.

Na tarefa de detecção de notas, o sistema de detecção de eventos sonoros foi muito eficiente. Uma consequência direta dessa precisão é a alta métrica de revocação observada no experimento 2. Caso o sistema de detecção de eventos não detectasse os momentos inicias das notas, as mesmas não seriam classificadas, aumentando a quantidade de falsos positivos, e diminuindo o valor de revocação.

O processo de geração de dados sintéticos demonstrou ser o suficiente para esse trabalho, porém diversas melhorias poderiam ser exploradas, como o condicionamento de notas a uma escala musical específica, e a menor chance de apinhamento de notas. Além disso, existem trabalhos que compõe melodias a partir de acordes pré definidos, essas

melhorias gerariam sinais que poderiam ser usados para testar um modelo de linguagem musical por exemplo.

Observa-se também que existem erros nos tempos iniciais das notas detectadas, esses erros podem ser causados por dois principais fatores, um erro na detecção exata da janela de início do evento e um erro que se acumula com o passar do sinal, gerado pela falha na detecção do ritmo da música em batidas por minuto.

6 Conclusão e Trabalhos Futuros

Infelizmente a performance alcançada no sistema proposto não foi alta o suficiente para o nível de aplicação comercial, porém demonstrou que ainda existem importantes desafios para serem vencidos, e quais são estes. Um exemplo são as melhorias citadas anteriormente na tarefa de separação de fontes sonoras, que caso tivesse ocorrido de foram perfeita como simulado no segundo experimento teria mais que duplicado a métrica pontuação F da solução proposta.

Melhorias futuras podem ser realizadas em todos os diferentes módulos do sistema, para que realizem suas tarefas de forma mais precisa, diminuindo a cascata de erros. É pouco esperado que exista mudança na estrutura geral da solução, que se demonstrou viável para a solução do problema. Além disso, melhorias podem ser realizadas no processo de geração de dados, para que diferentes situações possam ser especificamente geradas, como por exemplo o condicionamento da notas geradas a alguma escala musical, implementação de harmonias pré definidas ou randomicamente geradas, entre outras.

As melhorias no processo de gerar dados podem ser muito úteis para o treinamento e teste de modelos para as mais diferentes tarefas de MIR, como detecção de frequências fundamentais, separação de fontes sonoras, detecção de eventos sonoros e modelagem de linguagem musical. Além disso, a ferramenta poderia ser capaz de melhorar o acesso a datasets acessíveis e facilmente reproduzíveis. Por fim, alguma ferramenta gráfica também pode ser futuramente implementada para transformar a representação simbólica de nível computacional em nível de leitura humana, como partitura, cifras, entre outras.

Referências

ABU-MOSTAFA, Yaser S; MAGDON-ISMAIL, Malik. Lin, Hsuan-Tien. **Learning from Data: A Short Course. AMLBook. com**, 2012.

BENETOS, Emmanouil *et al.* Automatic music transcription: An overview. **IEEE Signal Processing Magazine**, IEEE, v. 36, n. 1, p. 20–30, 2018.

CAMACHO, Arturo; HARRIS, John G. A sawtooth waveform inspired pitch estimator for speech and music. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 124, n. 3, p. 1638–1652, 2008.

CHERKASSKY, Vladimir; MULIER, Filip M. **Learning from data: concepts, theory, and methods**. [S.l.]: John Wiley & Sons, 2007.

CHEUK, Kin Wai *et al.* The Effect of Spectrogram Reconstruction on Automatic Music Transcription: An Alternative Approach to Improve Transcription Accuracy. In: IEEE. 2020 25th International Conference on Pattern Recognition (ICPR). [S.l.: s.n.], 2021. P. 9091–9098.

DE CHEVEIGNÉ, Alain; KAWAHARA, Hideki. YIN, a fundamental frequency estimator for speech and music. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 111, n. 4, p. 1917–1930, 2002.

EMIYA, Valentin; BADEAU, Roland; DAVID, Bertrand. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 18, n. 6, p. 1643–1654, 2009.

GOWRISHANKAR, BS; BHAJANTRI, Nagappa U. An exhaustive review of automatic music transcription techniques: Survey of music transcription techniques. In: IEEE. 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). [S.l.: s.n.], 2016. P. 140–152.

HAWTHORNE, Curtis *et al.* Enabling factorized piano music modeling and generation with the MAESTRO dataset. **arXiv preprint arXiv:1810.12247**, 2018.

HENNEQUIN, Romain *et al.* Spleeter: a fast and efficient music source separation tool with pre-trained models. **Journal of Open Source Software**, v. 5, n. 50, p. 2154, 2020.

ISO 226:2003 (en). Acoustics — Normal equal-loudness-level contours. Standard, International Organization for Standardization. Geneva, CH, 2003.

JANSSON, Andreas *et al.* Singing voice separation with deep u-net convolutional networks, 2017.

KIM, Jong Wook *et al.* Crepe: A convolutional representation for pitch estimation. In: IEEE. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2018. P. 161–165.

KNEES, Peter; SCHEDL, Markus. **Music similarity and retrieval: an introduction to audio-and web-based strategies**. [S.l.]: Springer, 2016. v. 9.

LACERDA, Osvaldo. **Compêndio de teoria elementar da música**. [S.l.: s.n.], 1967.

MAUCH, Matthias; DIXON, Simon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In: IEEE. 2014 ieee international conference on acoustics, speech and signal processing (icassp). [S.l.: s.n.], 2014. P. 659–663.

MCFEE, Brian *et al.* librosa: Audio and music signal analysis in python. In: CITESEER. PROCEEDINGS of the 14th python in science conference. [S.l.: s.n.], 2015. v. 8, p. 18–25.

MCLEOD, Andrew; STEEDMAN, Mark. Evaluating Automatic Polyphonic Music Transcription. In: ISMIR. [S.l.: s.n.], 2018. P. 42–49.

THICKSTUN, John; HARCHAOUI, Zaid; KAKADE, Sham. Learning features of music from scratch. **arXiv preprint arXiv:1611.09827**, 2016.

VACA, Kevin; GAJJAR, Archit; YANG, Xiaokun. Real-time automatic music transcription (AMT) with Zync FPGA. In: IEEE. 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). [S.l.: s.n.], 2019. P. 378–384.

VAPNIK, Vladimir. **The nature of statistical learning theory**. [S.l.]: Springer science & business media, 1999.

WU, Yu-Te; CHEN, Berlin; SU, Li. Multi-Instrument Automatic Music Transcription With Self-Attention-Based Instance Segmentation. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, v. 28, p. 2796–2809, 2020.

WU, Yu-Te; CHEN, Berlin; SU, Li. Polyphonic music transcription with semantic segmentation. In: IEEE. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2019. P. 166–170.

YCART, Adrien *et al.* Blending acoustic and language model predictions for automatic music transcription, 2019.