Pontifícia Universidade Católica do Paraná
Programa de Pós-Graduação em Informática

Transcrição automática de músicas para Sanfona utilizando Transfer Learning

Lucas Mrowskovsky Paim

Orientador

Prof. Dr. Carlos N. Silla Jr.

Co-orientador

Prof. Dr. Alceu de Souza Britto Jr.

Curitiba
2024

Pontifícia Universidade Católica do Paraná
Programa de Pós-Graduação em Informática

Transcrição automática de músicas para Sanfona utilizando Transfer Learning

Lucas Mrowskovsky Paim

Dissertação apresentada ao Programa de Pós-Graduação em Informática como requisito parcial para obtenção do título de Mestre em Informática.

Campo de Concentração: Ciência da Computação

Orientador: Prof. Dr. Carlos N. Silla Jr.

Co-orientador: Prof. Dr. Alceu de Souza Britto Jr.

Curitiba
2024

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Luci Eduarda Wielganczuk – CRB 9/1118

P143t
2024

Paim, Lucas Mrowskovsky
Transcrição automática de músicas para sanfona utilizando transfer learning / Lucas Mrowskovsky Paim ; orientador: Carlos N. Silla Jr. ; co-orientador: Alceu de Souza Brito Jr. – 2024.
93 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2024
Bibliografia: f. 90-93

1. Informática. 2. Aprendizado do computador. 3. Transcrição automática de música. 4. Música para sanfona. 5. Inteligência artificial. 6. Algoritmos. I. Silla Junior, Carlos Nascimento. II. Britto Júnior, Alceu de Souza. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD. 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

Curitiba, 18 de setembro de 2024.

67-2024

DECLARAÇÃO

Declaro para os devidos fins, que **Lucas Mrowskovsky Paim** defendeu a dissertação de Mestrado intitulada "**Transcrição Automática de Músicas para Sanfona Utilizando Transfer Learning"**, na área de concentração Ciência da Computação no dia 24 de abril de 2024, no qual foi aprovado.

Declaro ainda, que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade firmo a presente declaração.

Documento assinado digitalmente
gov.br
EMERSON CABRERA PARAISO
Data: 18/09/2024 14:36:16-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Emerson Cabrera Paraiso
Coordenador do Programa de Pós-Graduação em Informática

Dedico este trabalho aos meus pais Rudnei e Edna e a minha noiva Thaís

Um homem aponta o céu.
O tolo olha o dedo, O sábio vê a lua.

Resumo

A transcrição automática de músicas é um campo que ainda necessita de muita pesquisa, pois ainda não há um algoritmo no estado da arte capaz de realizar uma transcrição completa de uma música, considerando a separação de vozes, detecção do tom da música, múltiplos instrumentos, etc. Um dos principais benefícios para alunos que estão iniciando no Sanfonaé que consigam ter uma melhor compreensão de como as duas mãos interagem na música. Este trabalho discute a aplicação da técnica transferência de conhecimento indutivo por transferência de representação de características de uma rede proposta inicialmente pelo grupo magenta para piano e seu desempenho aplicado ao Sanfona, assim como o impacto do uso de diferentes timbres (registros), a junção do baixo e o teclado tem sobre a transcrição e o impacto de se congelar diferentes camadas convolucionais durante o treinamento e a perca de conhecimento do problema original. Também foi criada a primeira base de dados de Sanfonafocada em Transcrição automática de músicas (AMT). Os resultados dos experimentos demonstraram uma melhora média na transcrição de músicas de Sanfonade aproximadamente 6% de f0-score, um aumento médio de 18% de precision e uma perca média de 6% de recall.

Palavras-chave: Transcrição aumática de música #1, Sanfona#2, Transferência de aprendizado #3

Abstract

Automatic music transcription is a field that still requires a lot of research, as there is currently no state-of-the-art algorithm capable of complete music transcription, considering voice separation, music key detection, multiple instruments, etc. One of the primary benefits for accordion students is to gain a better understanding of how both hands interact in music. This work discusses the application of the technique of inductive knowledge transfer through feature representation transfer from a network initially proposed by the Magenta group for the piano and its performance when applied to the accordion. It also explores the impact of using different timbres (registers), the combination of bass and keyboard on transcription, and the consequences of freezing different convolutional layers during training and losing original problem knowledge. The first accordion-focused Music Transcription Automatic (AMT) database was also created. The results of the experiments showed an average improvement in accordion music transcription of approximately 6% in F0-score, an average increase of 18% in precision, and an average loss of 6% in recall.

Key-words: Automatic Music Transcription #1, Accordion #2, Transfer-learning #3

Agradecimentos

Aos meus pais por todo apoio e paciência durante todos esses anos.

A minha noiva Thaís por não só compreender este momento de minha vida como por me ajudar revisando este documento e dando dicas de como melhorá-lo.

Ao meu orientador, prof. Dr. Carlos Silla, que acompanha a evolução deste trabalho desde a especialização, por sua paciência e incentivo para que eu permanecesse no mestrado apesar de diversos momentos conturbados de minha vida profissional.

Ao prof. Dr. Alceu e co-orientador deste trabalho, pelo voto de confiança durante a qualificação e por idealizar o novo protocolo de pesquisa deste trabalho.

Ao prof. Dr. Jean pelo voto de confiança e pelos diversos comentários valiosos durante a qualificação que me mostraram diversos pontos de melhoria para este trabalho.

Ao meu gestor Mahyar McDonald ("Mac") pela compreensão principalmente nesta reta final, me proporcionando a possibilidade de focar alguns dias durante a semana para a execução deste trabalho.

A prof. Marina Camargo do conservatório de MPB de Curitiba, por fornecer a gravações e a partituras que compõe parte importante do FoleDataSet.

À Pontifícia Universidade Católica do Paraná e toda sua equipe por me proporcionar esta oportunidade.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Sumário

Lista de Figuras

Lista de Símbolos

β Nome dado a primeira camada oculta da figura 3.4.

Δ Nome dado a segunda camada oculta da figura 3.4.

σ Nome dado a terceira camada oculta da figura 3.4.

δ Nome dado a última camada oculta da figura 3.4.

$\overleftarrow{\Delta}$ Representa todas as camadas anteriores, incluindo a camada Δ.

$\overrightarrow{\sigma}$ Representa todas as camadas posteriores, incluindo a camada σ.

λ Comprimento de onda.

A Amplitude de uma onda

f Frequência de uma onda

$f(\cdot)$ Função preditiva

$\mathcal{X}$ Espaço de características

$\mathcal{D}$ Domínio

$\mathcal{Y}$ Classes de um domínio

$\mathcal{T}$ Objetivo da função preditiva

Lista de Tabelas

Lista de Acrônimos

AMT Transcrição automática de músicas

BPM Batidas por minuto

FFT Fast fourier transform

Hz Hertz

midi Interface digital de instrumentos musicais

STFT Transformada curta de Fourier

wav waveform audio format

1

Introdução

A Transcrição automática de músicas (AMT) é o nome dado a um conjunto de algoritmos que visam converter música acústica em notação musical, como: partituras, midis, musicxml etc. (BENETOS et al., 2019).

Algoritmos de AMT facilitam que músicos que não possuem ouvido absoluto tenham acesso às transcrições musicais de forma facilitada e também que músicos iniciantes entendam melhor a relação entre notas e baixos da Sanfona, facilitando a compreensão de como as duas mãos interagem neste instrumento. (MACIEJEWSKI; LUKASIK, 2013) também citam que parte do interesse de músicos na transcrição automática é devido às improvisações de Jazz, pois muitas vezes os músicos não lembram exatamente quais notas tocaram durante uma improvisação.

Um sistema de AMT completo, segundo (BENETOS et al., 2019) ainda é considerado um desafio e um problema em aberto na literatura, principalmente para músicas que contenham múltiplas notas sendo tocadas simultaneamente (multi-pitch estimation) e múltiplos instrumentos, ou seja, ao fornecer um áudio de uma música clássica interpretada por uma orquestra, um sistema de AMT completo deve além de reconhecer todas as notas, mesmo que estejam sendo tocadas ao mesmo tempo, conseguir realizar a separação de instrumentos, como, por exemplo: primeiro violino, segundo violino, piano (mão esquerda e direita), etc.

Ainda de acordo com (BENETOS et al., 2019), um sistema de AMT completo deve:

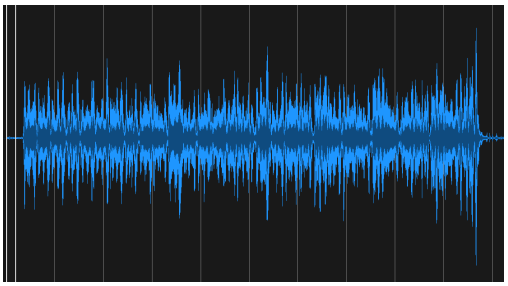
1. Reconhecer múltiplas notas simultâneas (multi-pitch estimation);

2. Detectar o início e a duração de uma nota (onset and offset detection);
3. Reconhecer o instrumento;
4. Reconhecer o ritmo;
5. Detectar a interpretação / expressividade do músico;
6. Transcrever o áudio em notação musical;

Em geral, o modus operandi de um sistema de AMT baseado em aprendizagem de máquina é: um arquivo de áudio como entrada (figura 1.1a), a partir disso um espectrograma é calculado (figura 1.1b) e a saída do algoritmo pode ser tanto uma partitura (figura 1.1c), arquivo Midi como na figura 1.1d ou um arquivo no formato MusicXML.

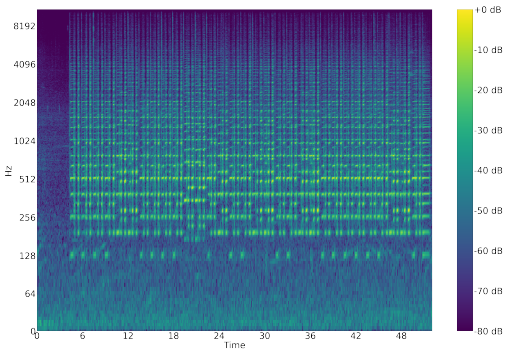
Figura 1.1: Componentes de um sistema AMT

(a) Primeira Valsa - WaveForm - 120bpm.



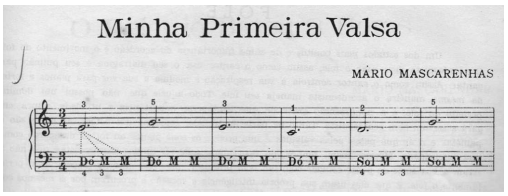
Fonte: Autoria Própria

(b) Short Time Fourier Transform - Espectrograma - Primeira Valsa - 120bpm.



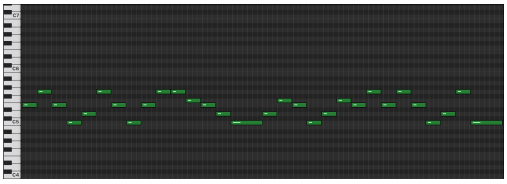
Fonte: Autoria Própria

(c) Primeira Valsa - Partitura



Fonte: (MASCARENHAS, 2006)

(d) Primeira Valsa - Apenas Teclado - Midi - 120bpm.



Fonte: Autoria Própria

Fonte: Autoria Própria

1.1 Justificativa

A Sanfona é um instrumento que tem grande influência na música folclórica e popular de países europeus. No território brasileiro ele é amplamente utilizado na região sul, principalmente no Rio Grande Do Sul e Santa Catarina, além do nordeste brasileiro. Diversos acordeonistas brasileiros merecem destaque, como: Luiz Gonzaga, Dominguinhos, Edson Dutra, Chiquinho do acordeão, Tio Bilia entre tantos outros. (CAMARGO, 2018)

Embora faça parte da cultura brasileira, o aprendizado da Sanfona ainda apresenta vários desafios. Na questão técnica do aprendizado (PAIVA, 2014), explica que uma das principais dificuldades decorre da complexidade de coordenação motora exigida, uma vez que mão esquerda frequentemente executa a base rítmica, a mão direita executa a melodia. Outra dificuldade encontrada é a grande falta de material didático de qualidade e partituras quando comparado a outros instrumentos. Estes motivos por muitas vezes causam a desistência do aprendizado da Sanfona.

1.2 Objetivos

O principal objetivo deste trabalho é investigar o uso de transferência de aprendizado de um modelo de AMT pré-treinado de piano para transcrições de Sanfona.

1.3 Hipótese

A abordagem proposta utilizando transferência de aprendizado terá um desempenho superior ao modelo pré-treinado do piano e ao modelo treinado a partir de pesos aleatórios.

1.4 Contribuições

A pesquisa desenvolvida nesse trabalho tem potencial de auxiliar na elaboração de partituras para a Sanfona.

Outra importante contribuição deste trabalho é a criação do primeiro dataset de músicas de Sanfona focado em AMT.

1.5 Estrutura do Documento

A estrutura deste documento está organizada da seguinte forma: o capítulo dois apresenta ao leitor uma introdução aos conceitos musicais necessários para o entendimento do restante deste trabalho. No capítulo três é aprofundado sobre o conceito de transferência de aprendizado assim como suas categorizações e técnicas. O capítulo quatro, por sua vez, apresenta o estado da arte. O capítulo cinco apresenta o método utilizado neste trabalho, demonstrando como é feito o pré-processamento do áudio e arquivos Interface digital de instrumentos musicais (midi), a criação da base de dados e o método de avaliação e validação utilizados. O capítulo seis, por sua vez, apresenta os experimentos realizados e por fim o capítulo sete apresenta as conclusões do trabalho.

2

Sons e Conceitos Musicais

Para a física, uma onda é um sinal transmitido de um ponto a outro sem que haja transporte de matéria, existindo dois tipos principais de onda: a mecânica e a eletromagnética. A diferença entre ambos os tipos é que a onda mecânica necessita de meios materiais para se propagar, exemplos deste tipo de onda são: a onda do mar, ondas sonoras, etc. Enquanto a onda eletromagnética pode se propagar no vácuo, como, por exemplo: a luz, raios-x, etc. (HALLIDAY; RESNICK; WALKER, 2008).

Se a direção de uma onda é a mesma direção da vibração do objeto que emitiu esta onda, é chamada de onda transversal e quando a direção da onda é perpendicular a vibração a onda é chamada de longitudinal (HALLIDAY; RESNICK; WALKER, 2008).

Som é definido como algo que possa ser ouvido (CAMBRIDGE-WEBSTER, 2023). Para a física som é definido como uma onda mecânica longitudinal, portanto possuí propriedades físicas, tais como: frequência (f), duração, amplitude (A), comprimento de onda (λ), etc. Na figura 2.1 são apresentados exemplos de ondas de baixa e alta frequência. (CARTER, 2018), (HALLIDAY; RESNICK; WALKER, 2008).

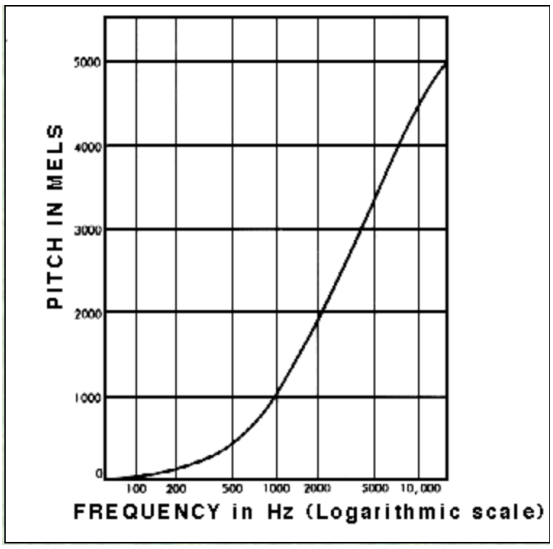
A unidade de medida de frequência é Hertz (Hz), que representa a quantidade de vezes que uma onda se repete por segundo. Para ondas sonoras a frequência é o que define o tom. Um ser humano consegue ouvir frequências entre 20Hz e 20.000Hz chamado intervalo audível, quaisquer valores fora deste intervalo é inaudível e chamado de "infra/sub"-sônico para frequências inferiores e ultra-sônico para valores que ultrapassam o intervalo (CARTER, 2018).

Figura 2.1: Ondas

(a) Baixa frequência (b) Alta frequência

Fonte: Autoria Própria

Figura 2.2: Escala de mel



Fonte: (STANFORD-WEBMASTER, s.d.)

Mesmo no intervalo audível por humanos não conseguimos distinguir frequências linearmente, ou seja, a maioria das pessoas consegue diferenciar facilmente um som de 100Hz e 200Hz, mas o mesmo não é verdade para 1000Hz e 1100Hz. Foi então que em 1937 Stevens, Volkmann e Newmann propuseram uma nova unidade de medida que iguala o intervalo, nomeada de Escala de Mel, demonstrada na figura 2.2. (PEDERSEN, 1965) (STANFORD-WEBMASTER, s.d.).

A amplitude (A) de uma onda sonora determina a altura de um som, quanto maior a amplitude mais alto será este som. (CARTER, 2018)

Tabela 2.1: Escala pitagórica

Inglês	C	D	E	F	G	A	B
América Latina	dó	ré	mi	fá	sol	lá	sí
Francês	ut	rè	mi	fà	sol	là	si

2.1 Notas musicais

Uma nota musical é apenas uma frequência que foi nomeada, por exemplo, a frequência 110Hz é associada a nota LÁ (A). (CARTER, 2018)

A origem da escala musical remonta ao matemático grego Pitágoras que, usando um monocórdio com um suporte móvel entre as extremidades fixas da corda vibrante, identificou as relações entre as frequências como os fatores preponderantes para a consonância dos sons. (GOTO, 2009)

A escala musical ocidental, ou escala diatônica, contém sete notas musicais como demonstrado na tabela 2.1. Estas notas também são chamadas de notas naturais.

A nomenclatura destas notas veio séculos depois de Pitágoras. O monge italiano Guido d'Arezzo (992-1050) vendo a necessidade de padronizar e simplificar a transcrição musical, idealizou o sistema de nomeação das notas, a qual é utilizada até hoje, utilizando as iniciais do "Hino a São João Batista". (MED, 1996)

A figura 2.3 demonstra todas as doze notas da música ocidental, sendo as sete notas naturais e as cinco notas conhecidas como "acidentes" e suas respectivas frequências. As notas com acidentes são as notas descritas com: # (sustenido) ou *b* (bemol). (ECHEVERRI; RODRIGUEZ; GARDUñO-APARICIO, 2018)

Cada nota tem sua própria frequência, mas existem poucos "nomes" ao se observar um instrumento, por exemplo, o piano, nota-se que há mais de doze teclas. Esses nomes são cíclicos, ou seja, assim que acabam eles se repetem. Uma oitava é apenas a distância de uma determinada nota até que esta se repita. (ECHEVERRI; RODRIGUEZ; GARDUñO-APARICIO, 2018) (CARTER, 2018)

O timbre é uma característica do som que nos permite diferenciar o som de mesma frequência e altura de um piano do som produzido por um violino. Na onda sonora o timbre é definido pelo formato da onda. Cada instrumento tem seu timbre característico, sendo o que nos permite diferenciá-los, uma flauta, irá produzir uma onda senoidal quase perfeita, enquanto o violino irá produzir uma onda mais complexa. (SILVA, 2023).

Figura 2.3: Frequências e notas

	Octaves				
Notes (Hertz)	**1**	**2**	**3**	**4**	**5**
C	32	65	130	261	523
C#	34	69	138	277	554
D	36	73	146	293	587
D#	38	77	155	311	622
E	41	82	164	329	659
F	43	87	174	349	698
F#	46	92	185	369	739
G	49	98	196	392	784
G#	52	104	208	415	830
A	55	110	220	440	880
A#	58	116	233	466	932
B	61	123	246	493	987

Fonte: (ECHEVERRI; RODRIGUEZ; GARDUñO-APARICIO, 2018)

Tabela 2.2: Escala de Dó (C) Maior

Tom	**Tom**	**Semitom**	**Tom**	**Tom**	**Tom**	**Semitom**	**Semitom**
C	**D**	**E**	**F**	**G**	**A**	**B**	**C**

2.2 Escalas e Acordes

Uma escala é um intervalo de notas bem definido. A escala maior é baseada na sequência lógica da nota Dó (C), sendo: $C- > D- > E- > F- > G- > A- > B$. Como dito anteriormente e demonstrado na figura 2.3, existem notas dentro dessa sequência que são chamados de acidentes, então podemos dizer que a sequência é definida por: tom-tom-semitom-tom-tom-tom-semitom-semitom como demonstrado na tabela 2.2. (CARTER, 2018)

Essencialmente um acorde é a combinação de duas ou mais notas, para escolher quais as notas da formação de um acorde estes são escolhidos com base em intervalos.

Os acordes maiores são montados em intervalos de terças, ou seja, a partir de uma nota da escala conta-se três notas a frente. Este padrão também é conhecimento como intervalo $1 - 3 - 5$, por exemplo, para montar o acorde de Dó (C) maior, se utiliza três notas: $C - E - G$ como demonstrado na tabela 2.3. (CARTER, 2018)

Existem acordes mais complexos e outras escalas, mas este trabalho não irá se aprofundar nestes conceitos.

Tabela 2.3: Acorde de Dó maior

C	E	G
1	3	5

Figura 2.4: Sheng

Fonte: (WIKIPEDIA-WEBSTER, 2024)

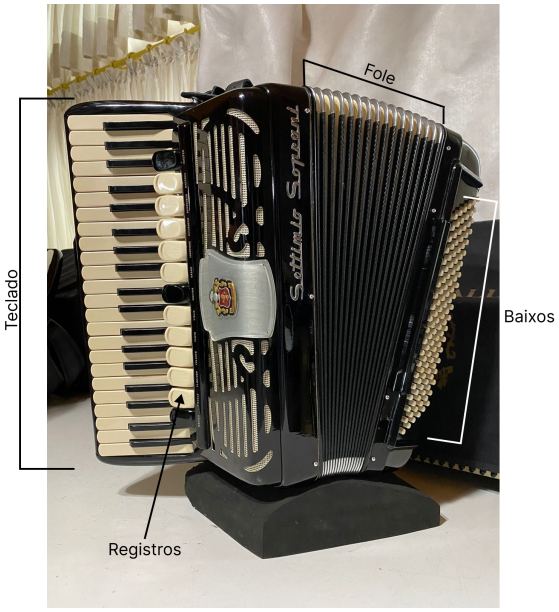
2.3 Sanfona

A origem da Sanfona é incerta, mas os princípios utilizados em sua fabricação são os mesmos de um instrumento chinês datado de 2700 a.C. chamado Sheng (Figura 2.4). A Sanfona funciona similarmente ao Sheng onde o som é produzido pelo ar sendo bombeado através das palhetas, porém na Sanfona o ar é bombeado pelo fole e não pela boca. Acredita-se que a Sanfona chegou ao Brasil no século XIX trazido pelos imigrantes italianos e posteriormente pelos alemães.(CAMARGO, 2018) (KLEBER, 2018)

A Sanfona Stradella moderno é composto por quatro partes principais: o teclado, registros, fole e os baixos como demonstrado na figura 2.5.

O teclado é utilizado para a melodia da música e fica na mão direita do

Figura 2.5: Partes da Sanfona



Fonte: Autoria Própria

acordeonista, os registros alteram o timbre do instrumento, os baixos ficam a mão esquerda sendo responsáveis pelo ritmo e o fole pressiona o ar pelas palhetas para produzir som.

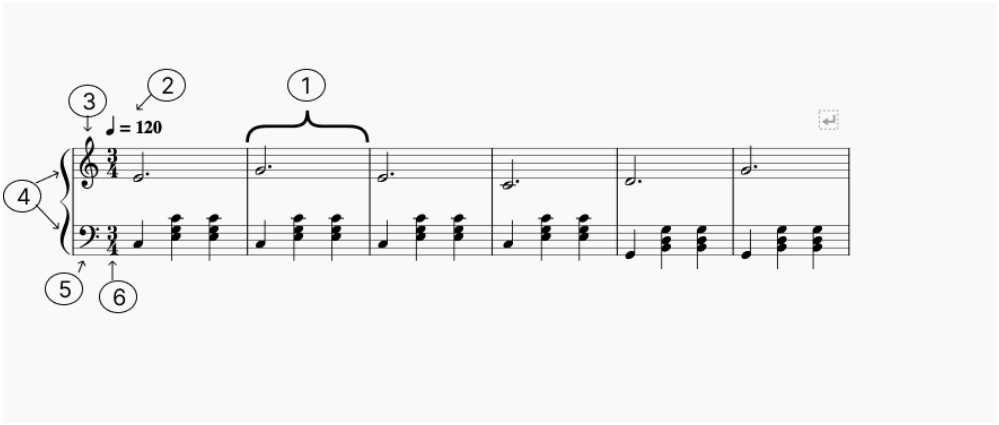
2.4 Partitura

A música também pode ser definida como notas estruturadas no tempo. A partitura é uma forma de representar esta estrutura. (SHATRI; FAZEKAS, 2020)

A figura 2.6 demonstra os principais componentes de uma partitura.

1. Compasso: representado pelas linhas verticais, indica a divisão da música em intervalos de tempos iguais. O tempo é definido pela quantidade de batidas por minuto definido pelos items 2 e 6.
2. Batidas por minuto (BPM)
3. Clave: A clave define a representação do tom, este simbolo demonstra a clave de Sol (G) isso significa que a segunda linha deste grupo é a nota sol e cada espaço e linha a partir dela representa um tom.
4. Vozes: Tanto quanto em partituras para Sanfona quanto para partituras de piano, esses dois grupos representam as notas tocadas por cada uma das mãos. O grupo de cima com a clave de SOL representa a mão direita (teclado da Sanfona) e a mão esquerda representa os baixos da Sanfona.

Figura 2.6: Partitura Primeira Valsa



Fonte: Adaptado de (MASCARENHAS, 2006)

5. Clave de Fá, assim como a clave de Sol (G) define o tom da segunda linha deste grupo.
6. Fórmula do compasso: O numerador define quantos pulsos o compasso tem e o denominador indica a figura que irá definir a duração do pulso no compasso.

2.5 Midi

Introduzido inicialmente em janeiro de 1983 na associação nacional de músicos mercantes, o formato midi 1.0 substituiu os primeiros projetos de uma interface para um sintetizador universal. Uma curiosidade sobre a versão 1.0 é que não houve envolvimento da academia ou de alguma comunidade de teoria musical, sendo apenas envolvidos posteriormente para resolver as lacunas deixadas na primeira especificação do formato. (MAZZOLA et al., 2018)

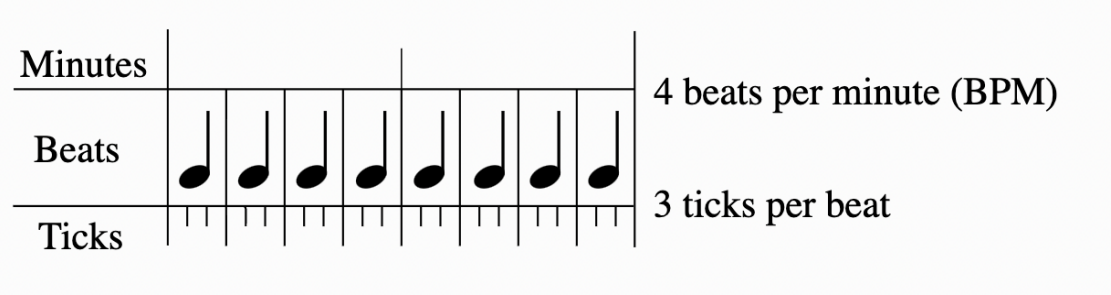
O formato midi é um arquivo baseado em sequência de eventos que devem ser interpretados pelo computador para a reprodução de uma música, os principais eventos deste arquivo utilizados neste trabalho são: *set_tempo*, *note_on* e *note_off*.

Segundo a documentação da biblioteca (MIDO, 2021) o BPM em um arquivo midi é definido em batidas *beats* e *ticks*.

Cada mensagem no arquivo midi informa quantos *ticks* passaram desde a última mensagem, pela propriedade "time" observável na figura 2.7 .

A figura 2.8 demonstra alguns exemplos de mensagens midi da música missioneiro (mão direita / teclado).

Figura 2.7: Exemplo de Eventos Midi



Fonte: (MIDO, 2021)

Figura 2.8: Exemplo de Eventos Midi

```
MidiTrack([
  MetaMessage('set_tempo', tempo=681818, time=0),
  Message('program_change', channel=0, program=0, time=0),
  MetaMessage('key_signature', key='G', time=0),
  MetaMessage('time_signature', numerator=2, denominator=4, clocks_per_click=24, notated_32nd_notes_per_beat=8, time=
0),
  Message('note_on', channel=0, note=43, velocity=80, time=30720),
  Message('note_off', channel=0, note=43, velocity=0, time=10912),
  Message('note_on', channel=0, note=55, velocity=80, time=608),
  Message('note_on', channel=0, note=59, velocity=80, time=0),
  Message('note_on', channel=0, note=62, velocity=80, time=0),
  Message('note_off', channel=0, note=55, velocity=0, time=3616),
  Message('note_off', channel=0, note=59, velocity=0, time=0),
  Message('note_off', channel=0, note=62, velocity=0, time=0),
  Message('note_on', channel=0, note=47, velocity=80, time=224),
  Message('note_off', channel=0, note=47, velocity=0, time=7264),
  Message('note_on', channel=0, note=55, velocity=80, time=416),
  Message('note_on', channel=0, note=59, velocity=80, time=0),
  Message('note_on', channel=0, note=62, velocity=80, time=0),
```

Fonte: Autoria Própria

3

Transferência de aprendizado

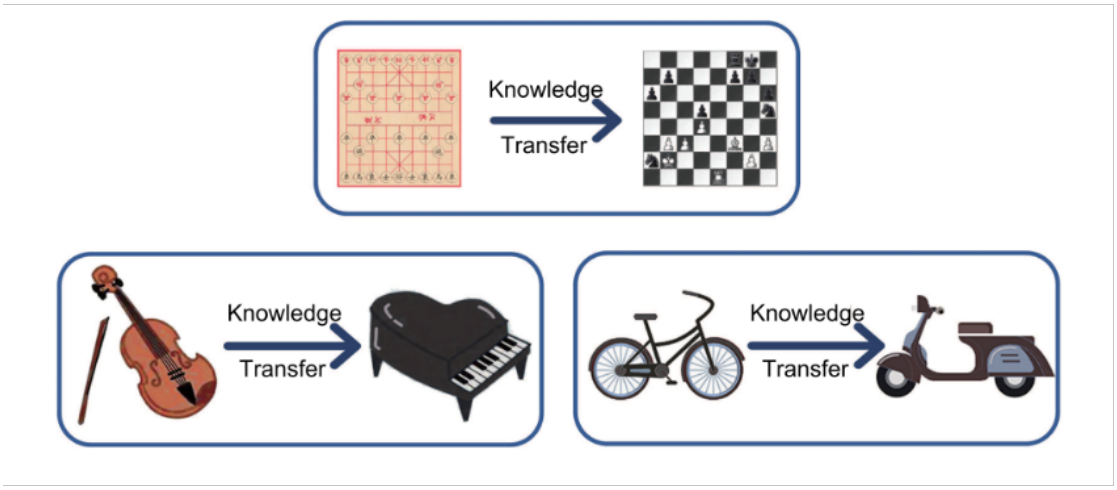
O treinamento de redes neurais densas exige um volume considerável de dados e poder computacional para atingir o resultado desejado, mas a criação e catalogação desses dados é um processo que leva muitas horas de laboratório e, portanto, tem um custo muito elevado. Inspirado na habilidade humana de transferir conhecimento entre domínios, como ao aprender um primeiro instrumento se torna mais fácil aprender um segundo, as técnicas de transferência de aprendizado consistem em aplicar o mesmo conceito em algoritmos de aprendizado de máquina. A figura 3.1 demonstra alguns exemplos intuitivos de transferência de aprendizado. (GÉRON, 2019), (WEI et al., 2018) e (ZHUANG et al., 2020)

Para este trabalho é considerado a seguinte terminologia:

1. Espaço de características: Dados apresentados ao algoritmo, normalmente representado pela letra $\mathcal{X}$, onde $\mathcal{X} = \{x_1, x_2 ... x_n\}$;
2. Domínio: É a combinação do espaço de características e a probabilidade de distribuição marginal $P(\mathcal{X})$, normalmente representado pela letra $\mathcal{D}$, onde $\mathcal{D} = \{\mathcal{X}, P(\mathcal{X})\}$
3. Tarefa: Conjunto que representa a junção das classes $\mathcal{Y}$ e a função preditiva, normalmente representada pela letra $\mathcal{T}$, sendo $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

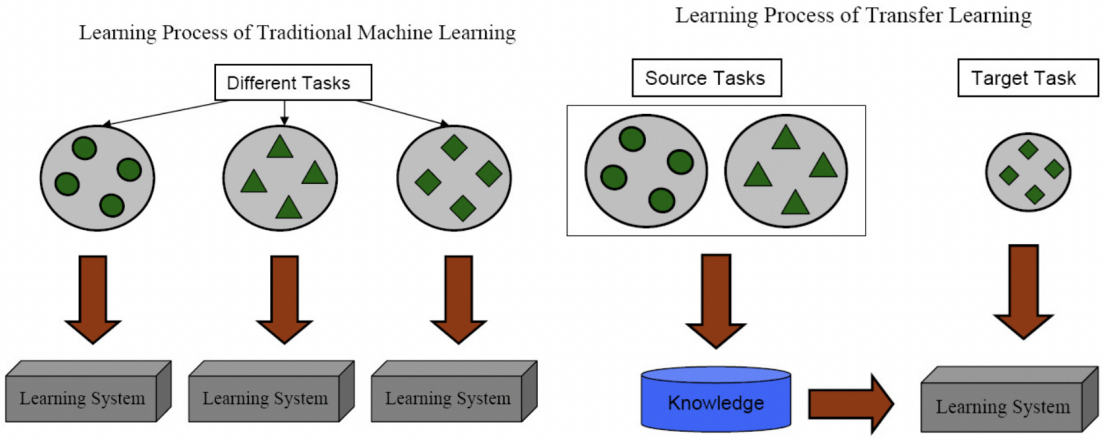
A figura 3.2 demonstra a diferença entre as técnicas convencionais de aprendizagem de máquina e as que utilizam transferência de aprendizagem, em suma a principal diferença é que em técnicas de transferência de aprendizagem os classificadores treinados para tarefas diferentes são re-utilizados para tarefas similares.

Figura 3.1: Exemplos intuitivos de transferência de aprendizado



Fonte: (ZHUANG et al., 2020)

Figura 3.2: Diferentes processos de aprendizagem.



(a) Traditional Machine Learning (b) Transfer Learning

Fonte: (PAN; YANG, 2010)

Definição de Transferência de aprendizado: Dado um domínio inicial $\mathcal{D}_s$ e uma tarefa inicial $\mathcal{T}_s$ e um domínio alvo $\mathcal{D}_t$ e uma tarefa alvo $\mathcal{T}_t$ a técnica de transferência de aprendizado tenta melhorar o aprendizado do classificador $f(\cdot)$ no $\mathcal{D}_t$ utilizando o conhecimento adquirido em $\mathcal{D}_s$ e $\mathcal{T}_s$ desde que $\mathcal{D}_s \neq \mathcal{D}_t$ ou $\mathcal{T}_s \neq \mathcal{T}_t$. (PAN; YANG, 2010) (Tradução pória)

Ao utilizar estas técnicas, pode-se evitar overfitting em datasets considerados

pequenos e com o custo computacional exponencialmente menor do que utilizando a abordagem convencional, pois não é necessário treinar o classificador do zero para atingir resultados satisfatórios. (GÉRON, 2019)

3.1 Categorização

A transferência de aprendizado pode ser categorizada em três grandes grupos: indutivo, transdutivo e não supervisionado, dependendo da relação entre o domínio e tarefas iniciais do algoritmo e do domínio e tarefa alvo. (PAN; YANG, 2010)

3.1.1 Indutivo

O aprendizado de máquina indutivo, também conhecido como aprendizado supervisionado, é quando a base de treinamento contém os rótulos desejados. Normalmente é utilizado em algoritmos de classificação como, por exemplo, classificação de objetos em imagens. (GÉRON, 2019) (MALLAWAARACHCHI, 2020)

A transferência de aprendizado indutivo, refere-se à quando a tarefa inicial e alvo diferem $\mathcal{T}_s \neq \mathcal{T}_t$, porém, são relacionadas independentemente se ambos os domínios são relacionados ou não. Um cenário onde isso pode ser aplicado é na transferência de aprendizado entre o reconhecimento de estrelas e o reconhecimento de galaxias, neste cenário o domínio pode até ser o mesmo, mas a tarefa apesar de diferir é relacionada. (PAN; YANG, 2010) (VILALTA et al., 2010)

Dependendo se a base de dados do domínio original está rotulada ou não, a transferência de aprendizado indutivo ainda pode ser classificada em dois outros grupos:

- Multi-task Learning: Quando os rótulos do domínio original estão disponíveis, e ambos os classificadores são treinados ao mesmo tempo.
- Self-taught Learning: Quando os rótulos do domínio original não estão disponíveis, mas o do domínio alvo estão.

Se o treinamento da tarefa original e da tarefa alvo são realizadas ao mesmo tempo, também pode ser chamado de transferência representacional, caso seja em momentos diferentes é chamado de transferência funcional. (VILALTA et al., 2010)

3.1.2 Transdutivo

A transferência de conhecimento é nomeada transdutiva quando há rótulos no domínio original, mas não há no domínio alvo, neste caso $\mathcal{D}_s \neq \mathcal{D}_t$ e $\mathcal{T}_s = \mathcal{T}_t$. (PAN; YANG, 2010)

3.1.3 Não supervisionado

O aprendizado de máquina não supervisionado é quando a base de dados não apresenta os rótulos, nesta técnica o objetivo do algoritmo pode ser realizar agrupamentos dos dados, redução de dimensionalidade, detecção de anomalias, etc. (GÉRON, 2019)

Neste caso não há rótulos tanto no domínio original quanto no domínio alvo. (PAN; YANG, 2010)

3.2 Técnicas

No estado da arte é possível encontrar diversas técnicas para a aplicação de transferência por aprendizado, estas técnicas podem ser classificadas em quatro grupos: (PAN; YANG, 2010)

1. Transferência por Instâncias
2. Representação de características
3. Transferência por parâmetros
4. Transferência de conhecimento relacionado.

3.2.1 Transferência por instâncias

Nesta técnica, parte dos dados do domínio original são utilizados durante o treinamento do novo classificador. (PAN; YANG, 2010) (HOSSAIN et al., 2018)

Ao aplicar esta técnica é necessário analisar se as instâncias utilizadas do domínio do original se adaptam ao novo domínio ou a transferência de aprendizado será dificultado por um problema chamado "transferência negativa" que é quando a transferência de aprendizado tem um efeito negativo no problema alvo. (PAN; YANG, 2010) (HOSSAIN et al., 2018)

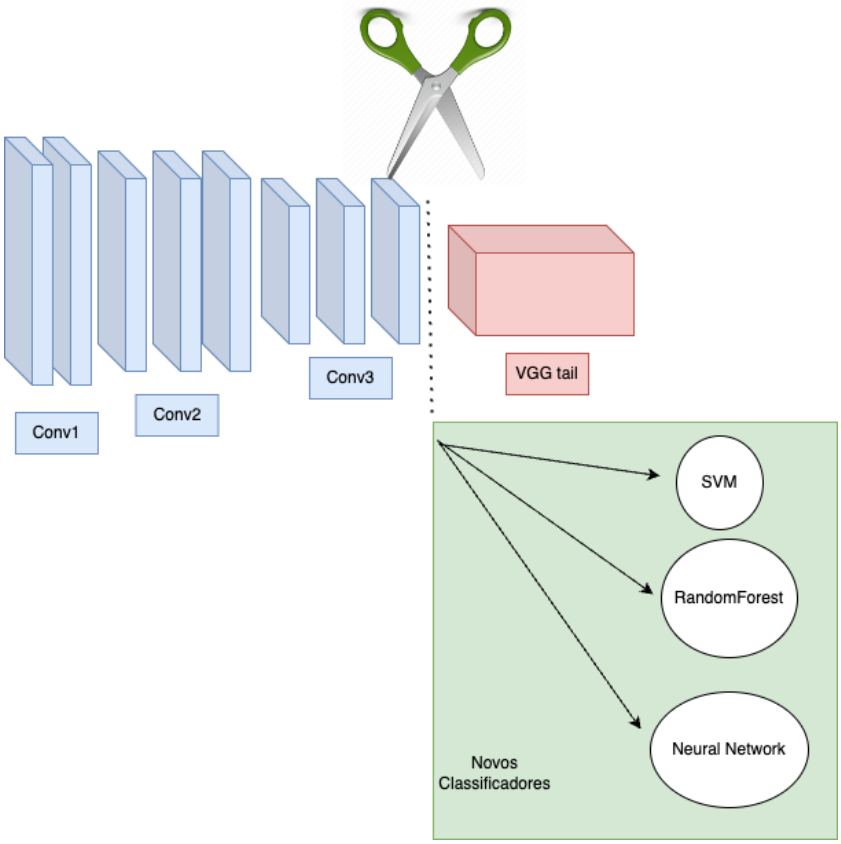
3.2.2 Representação de características

Esta técnica consiste em encontrar uma representação de características que seja aplicável aos dois problemas (PAN; YANG, 2010).

Quando aplicado em redes neurais, consiste em obter uma rede neural pré-treinada, congelar algumas camadas e treinar as demais. Algumas variações desta técnica podem substituir completamente parte da rede para uma nova topologia ou treinar outros classificadores com base na saída da rede neural, como demonstrado na figura 3.3 que representa a cauda da rede VGG sendo removida da rede e os pesos sendo utilizados para treinar outros classificadores. (GÉRON, 2019)

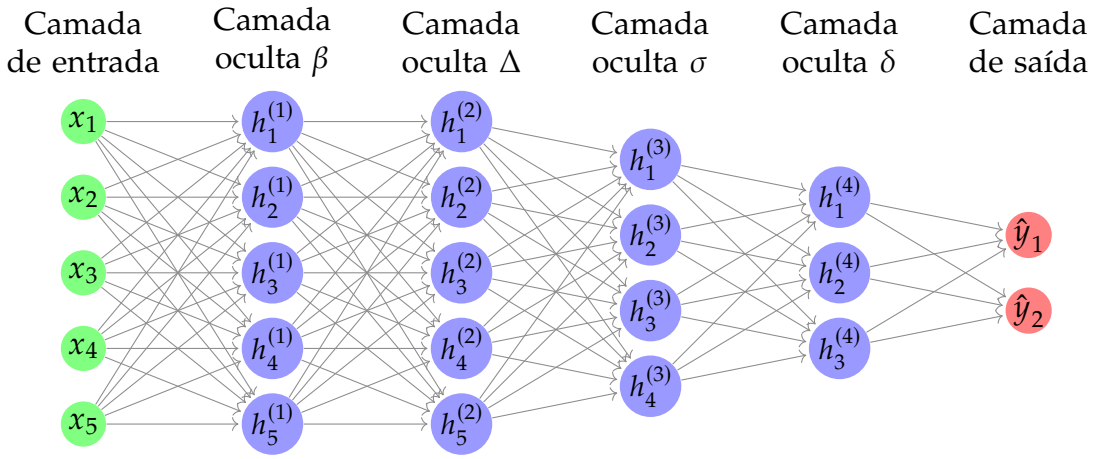
A figure 3.4 mostra um exemplo de uma rede neural, em outra variação desta técnica, é possível congelar apenas parte das camadas e treinar o restante da rede neural, por exemplo, pode-se congelar as camadas $\overleftarrow{\Delta}$ e treinar apenas as camadas $\overrightarrow{\sigma}$ para utilizarem o conhecimento adquirido em $\mathcal{T}_s$ em $\mathcal{T}_t$.

Figura 3.3: Transferência por características.



Fonte: Autoria Própria

Figura 3.4: Exemplo de rede neural



Fonte: Autoria Própria

3.2.3 Transferência de conhecimento relacionado

A técnica de transferência de conhecimento relacionado consiste em encontrar as relações ou regras relacionadas em ambos os domínios. Como, por exemplo, pode-se considerar que um professor exerce uma função no domínio acadêmico similar a um gestor no domínio industrial. (PAN; YANG, 2010) (BAHETI, 2021)

3.2.4 Transferência de parâmetros

A técnica de transferência por parâmetros, consiste em encontrar parâmetros dos classificadores que podem ser utilizados tanto para a tarefa original quanto para a tarefa alvo. (PAN; YANG, 2010)

4

Trabalhos relacionados

Para realizar este trabalho foi feita uma busca em bases de dados científicas por trabalhos na área de transcrição automática que utilizassem a Sanfona. Contudo, só foi encontrado um único rascunho de artigo, desenvolvido por (MACIEJEWSKI; LUKASIK, 2013).

No trabalho de (MACIEJEWSKI; LUKASIK, 2013), os autores discutem os problemas relacionados a este instrumento. Apesar de discorrerem sobre a utilização da técnica de multiplicação de matrizes não negativas, que não pertence ao escopo deste trabalho, vários pontos levantados foram relevantes a este estudo e que vão ser discutidos na seção 4.1.

Apesar de só ter sido encontrado o trabalho de (MACIEJEWSKI; LUKASIK, 2013) para a Sanfona, foi expandido a busca para trabalhos que utilizassem técnicas para o piano. Alguns dos trabalhos encontrados foram: (HAWTHORNE et al., 2018) demonstrado na seção 4.2, (BöCK; SCHEDL, 2012), (SIGTIA; BENETOS; DIXON, 2016), (EMIYA; BADEAU; DAVID, 2010) e (KELZ et al., 2016).

Os algoritmos de AMT, segundo (BENETOS et al., 2019) podem ser classificados em duas categorias:

1. *non-negative matrix factorization* (NMF)
2. Redes Neurais

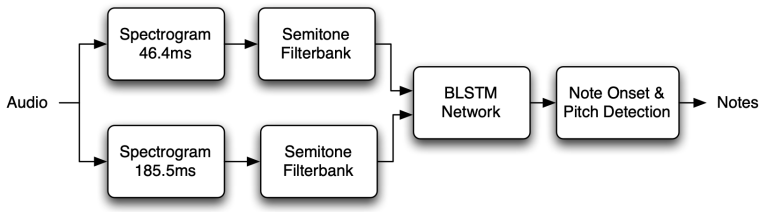
Considerando que o objetivo deste trabalho consiste em estudar técnicas de transferência de aprendizado, no restante desta seção serão apresentados apenas trabalhos que utilizam redes neurais.

Segundo (KELZ et al., 2016) algumas técnicas consistem na separação do problema em dois modelos diferentes, o modelo acústico e o modelo de linguagem musical ou em apenas um único modelo que realize as duas tarefas. Os autores demostraram o potencial de redes convolucionais para realizar a classificação acústica, demonstrando como estas redes podem classificar corretamente as notas presentes em uma música a partir de um espectrograma utilizando arquiteturas de redes como: DNN, ConvNet e AllConv utilizando o MAPS dataset (EMIYA et al., 2010).

Já (SIGTIA; BENETOS; DIXON, 2016) mencionam que o modelo de linguagem musical não está presente em diversas técnicas de AMT, devido à complexidade de se modelar as possibilidades das notas para técnicas polifônicas e também demonstram o ganho de acurácia ao se combinar os dois modelos (acústico+linguagem) em seu trabalho os autores propõe o uso da transformada de Q-Constant e obtiveram um f-measure de 74.45% no MAPS dataset.

O trabalho realizado por (BöCK; SCHEDL, 2012) demonstra a capacidade que camadas LSTM bidirecionais tem para realizar a transcrição. A topologia proposta é demonstrada na figura 4.1, onde dois espectrogramas de Fourier utilizando duas janelas diferentes são apresentadas para a rede. Esta abordagem obteve uma acurácia de 84% no MAPS dataset.

Figura 4.1: Topologia proposta por (BöCK; SCHEDL, 2012)



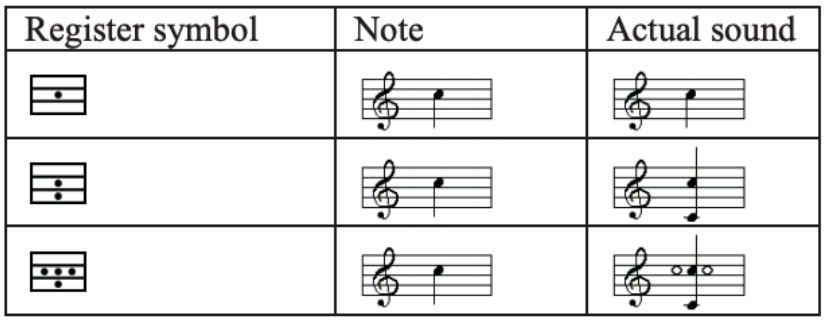
Fonte: (BöCK; SCHEDL, 2012)

4.1 AMT aplicado a Sanfona

O trabalho realizado por (MACIEJEWSKI; LUKASIK, 2013) demonstra alguns problemas relacionados a transcrição de músicas de Sanfona, muitos destes problemas ocorrem devido à natureza polifônica deste instrumento e suas

diferentes vozes que podem ser alterados com o uso dos registros, como demonstrado na figura 4.2 e explicado na seção 2.3.

Figura 4.2: Mudança de registros



Fonte: (MACIEJEWSKI; LUKASIK, 2013)

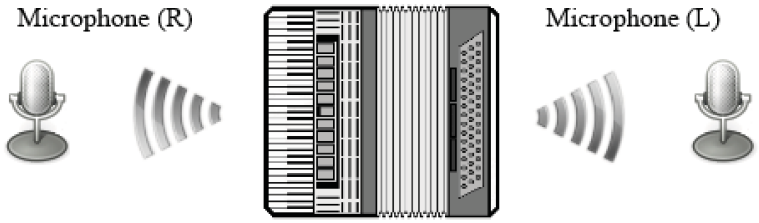
Para cada registro as vozes da Sanfona mudam, conforme a figura 4.2, no primeiro registro demonstrado, o som e nota tocada na Sanfona são os mesmos. No segundo registro demonstrado, com apenas uma nota sendo tocada na Sanfona o som produzido é a nota em conjunto com a mesma nota uma oitava abaixo, no último registro demonstrado além da oitava abaixo em conjunto, mais duas vezes com a mesma frequência da nota são tocadas ao mesmo tempo.

(MACIEJEWSKI; LUKASIK, 2013) também citam que tanto os baixos quanto o teclado da Sanfona produzem áudio com características muito próximas, tornando a etapa de sua separação bastante difícil, se não impossível.

A metodologia proposta por (MACIEJEWSKI; LUKASIK, 2013) consiste em gravar dois áudios, utilizando dois microfones em modo estéreo como demonstrado na figura 4.3 e depois usando um espectrograma da transformada de Fourier é aplicada um algoritmo de NMF.

Em relação aos experimentos realizados, conforme a interpretação de (MACIEJEWSKI; LUKASIK, 2013) os resultados foram satisfatórios para melodias simples. Porém, o trabalho não apresenta detalhadamente o seu protocolo experimental, ou seja, não é possível replicar o trabalho, pois não foram detalhadas quais melodias foram avaliadas, nem qual métrica foi utilizada.

Figura 4.3: Gravação em modo stereo.



Fonte: (MACIEJEWSKI; LUKASIK, 2013)

4.2 Onset and Frames - Google Brain

Dada a falta de trabalhos para Sanfona, expandiu-se a pesquisa para técnicas desenvolvidas para outros instrumentos, como o piano.

A rede com a melhor acurácia para piano atualmente no estado da arte é a rede desenvolvida por (HAWTHORNE et al., 2018). O modelo calcula o espectrograma de Mel_{log} e o utiliza de entrada para os modelos, esta rede foi inicialmente treinada utilizando o MAPS dataset (EMIYA et al., 2010) e futuramente melhorada utilizando um novo dataset chamado Maestro (HAWTHORNE et al., 2019).

A rede consiste em dois modelos principais, um modelo acústico responsável pela detecção das notas e o modelo de linguagem responsável por ajustar a duração da nota na série temporal. (HAWTHORNE et al., 2018)

Após análise do código foi constatado que apesar de o artigo apenas citar indiretamente, na verdade, a rede é composta por quatro regiões: (1) Onset: Modelo responsável por detectar o início das notas, (2) Offset: Modelo responsável por detectar o final de uma nota, (3) Frame: Modelo responsável por detectar o tempo intermediário entre o início e o fim das notas e (4) Velocity: Modelo responsável por detectar o caimento da energia da nota de modo a criar transcrições mais realistas, como demonstrado na figura 4.4. Cada um destes modelos tem um grupo convolucional representado na figura 4.5.

O pré-processamento do modelo ocorre como demonstrado no algoritmo 1, convertendo o arquivo midi em um formato chamado piano-roll. Para o treinamento do modelo foi utilizado a função de perda chamada *Binary Cross Entropy* ou *Log Loss*, definida pela equação 4.1 e como a saída da rede é de 88

neurônios, cada loss individual é somado como demonstrado pela equação 4.2.

Esta função de perda tem uma limitação que é feito apenas para classificações binárias para poder utilizar esta função de perda com a saída do algoritmo 1, que contém três diferentes classes (início/meio/fim) de uma nota. Sendo necessário treinar cada parte do modelo utilizando apenas uma destas classes, sendo assim, cada parte da rede $f(\cdot)$ recebe como entrada uma matriz binária para cada classe alvo.

O otimizador utilizado é o algoritmo de Adam com uma taxa de aprendizagem de 10^{-3}. [1]

$$bLoss = -(y \log(p) + (1 - y) \log(1 - p)) \tag{4.1}$$

$$nLoss = \sum_{i=1}^{n} bLoss_i \tag{4.2}$$

4.3 Datasets

Não foi encontrado nenhum dataset especificamente para a Sanfona, porém no estado da arte existem dois datasets amplamente utilizados para transcrição musical de piano, são eles o Maps dataset [2] proposto por (EMIYA et al., 2010) e o Maestro dataset [3] proposto por (HAWTHORNE et al., 2019).

4.3.1 Maps

Este dataset foi proposto em 2010, ele consiste em áudios gravados utilizando um piano conectado a uma saída midi e um microfone próximo, como demonstrado na figura 4.6. Neste dataset existem tanto músicas, escalas e acordes isolados sendo tocados no piano, totalizando cerca de $\approx 65hrs$ de gravação e $\approx 40GB$ de dados.

[1] A Sanfona não possui pedais de sustentação de notas, tornando o bloco de correção do final da nota irrelevante, porém ele é necessário para se treinar o modelo para transcrição de piano corretamente

[2] http://www.tsi.telecom-paristech.fr/aao/en/category/database/

[3] https://magenta.tensorflow.org/datasets/maestro/

Algoritmo 1 Conversão de midi em piano-roll

Require: $duracao\_audio > 0$
1: $FRAME \leftarrow 32$ ▷ Em milisegundos
2: $DELTA \leftarrow \lfloor \frac{duracao\_audio}{FRAME} \rfloor$
3: $N\_NOTES \leftarrow 88$ ▷ Número de possíveis notas do piano
4: $PIANO\_ROLL \leftarrow [0]_{N\_NOTESXDELTA}$
5: $tempo\_acumulado \leftarrow 0$
6: **for** `evento` *in* `eventos_midi` **do**
7: **if** $evento.type \notin [''note\_on'', ''note\_off'']$ **then**
8: **skip**
9: **end if**
10: $tempo\_acumulado \leftarrow tempo\_acumulado + event.time$
11: $posicao = tempo\_acumulado \times 1000$
12: $posicao = \lfloor \frac{posicao}{FRAME} \rfloor$
13: $nota = event.note - 21$
14: **if** $evento.type = ''note\_on''$ **and** $evento.velocity > 0$ **then**
15: $PIANO\_ROLL[nota, posicao] \leftarrow 1$
16: **else**
17: $pedal\_sustain = relevant\_sustain\_message(evento)$
18: **if** $pedal\_sustain$ **is not null then** ▷ irelevante para Sanfona
19: $posicao = pedal\_sustain.tempo\_final \times 1000$
20: $posicao = \lfloor \frac{posicao}{FRAME} \rfloor$
21: **end if**
22: $PIANO\_ROLL[nota, posicao] \leftarrow 3$
23: **end if**
24: # <omitido> iterar $PIANO\_ROLL$ e preencher as lacunas entre os valores 1 e 3 com o valor 2, ex: [0, 1, 0, 3, 0] -> [0, 1, 2, 3, 0].
25: **end for**

4.3.2 Maestro

Proposto inicialmente em 2019, este dataset teve basicamente a mesma metodologia do Maps dataset, com a diferença que ele não contém escalas e acordes sendo tocados isoladamente.

Como mencionado em (HAWTHORNE et al., 2019), para a criação deste dataset foi realizada uma parceria com uma competição internacional de pianistas, onde um estúdio similar ao apresentado na figura 4.6b foi utilizado para realizar a gravação dos áudios e seus respectivos arquivos midi.

Este dataset em sua primeira versão continha $\approx 172hrs$ de gravação e $\approx 121.8GB$ de dados, atualmente este dataset está na sua versão 3.0 e contém $\approx 198.7hrs$ de gravação e $\approx 120.2GB$ de dados.

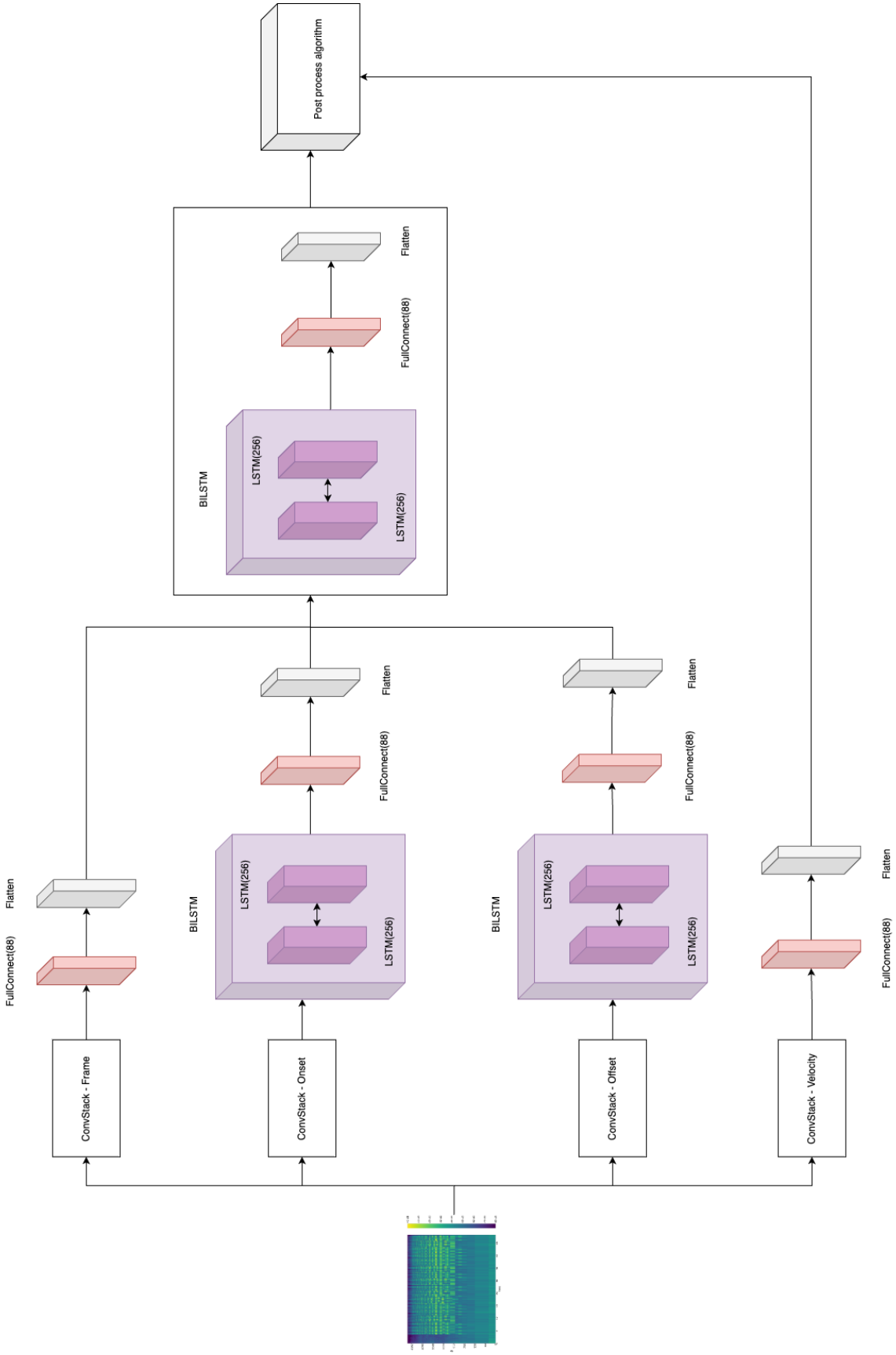
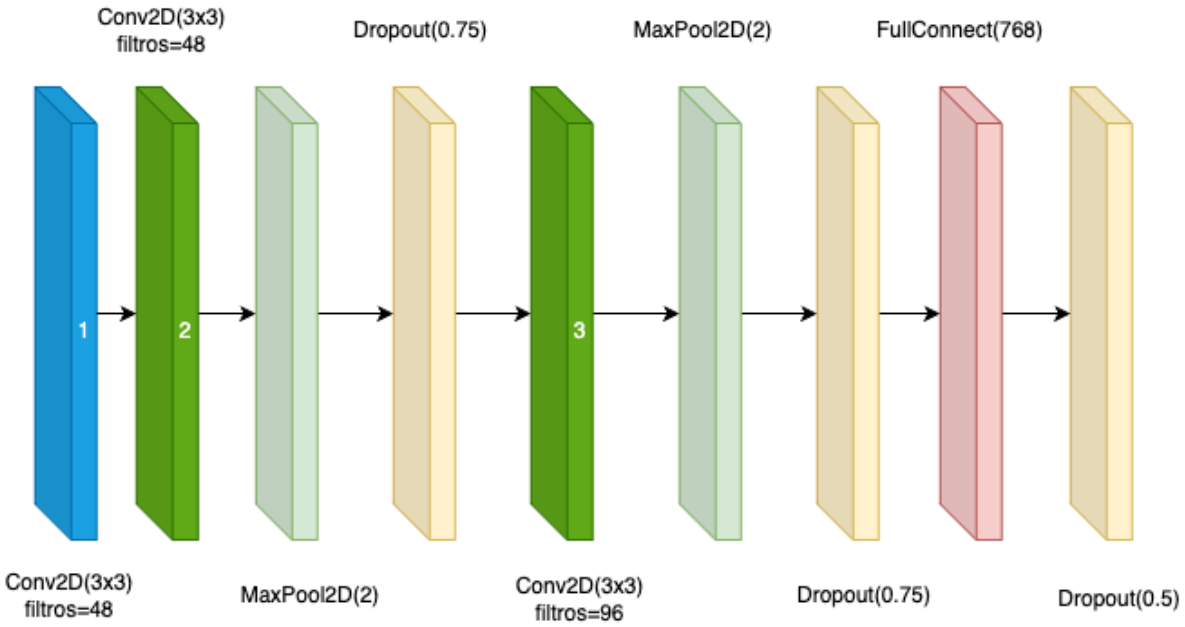


Figura 4.4: Arquitetura da Rede Onset And Frames

Fonte: Autoria Própria

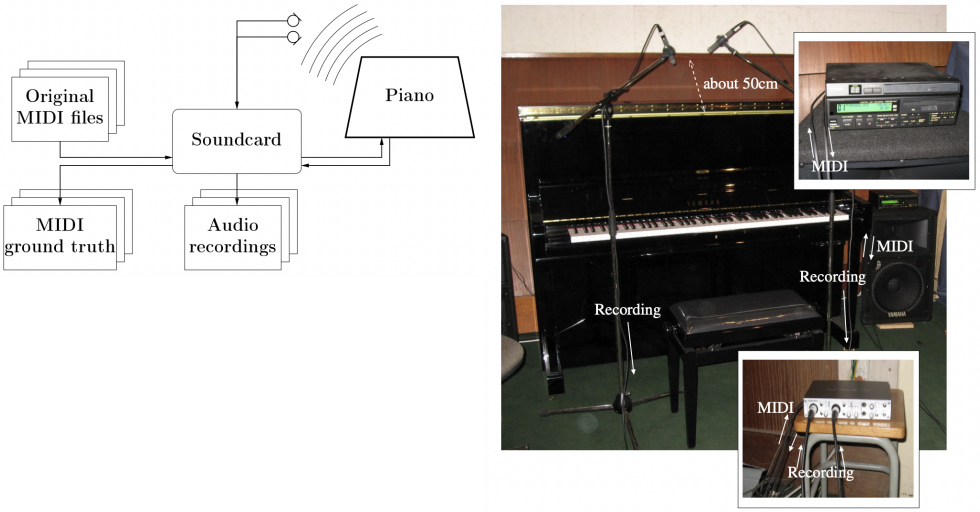
Figura 4.5: Região Convolucional da rede Onset And Frames



Fonte: Autoria Própria

Figura 4.6: Estúdio utilizado para a criação do maps dataset

(a) Diagrama de blocos (b) Estúdio de gravação do dataset



Fonte: (EMIYA et al., 2010)

5

Método

5.1 Análise Exploratória

Para uma análise preliminar foi escolhida a música Primeira Valsa publicada no método de aprendizado de Sanfona de Mário Mascarenhas com a interpretação realizada com o metrônomo a 120 bpm e configurado em 3/4, pois é a primeira música apresentada pelo método de ensino Mário Mascarenhas (MASCARENHAS, 2006). [1]

Utilizando o software Sonic Visualiser (CANNAM; LANDONE; SANDLER, 2010) para visualizar o espectrograma da transformada de Fourier STFT utilizando uma janela de tamanho 2^{12} e aumentar sua escala no eixo x, é possível notar o padrão das notas e dos baixos da Sanfona, como demonstrado na figura 5.1. Na figura 5.1b é possível visualizar os 3 primeiros compassos da peça, onde os círculos em vermelho demonstram respectivamente as notas MI-SOL-MI no teclado da Sanfona e os círculos verdes demonstram o fundamental SOL dos baixos e o amarelo demonstra o acorde de SOL maior também nos baixos.

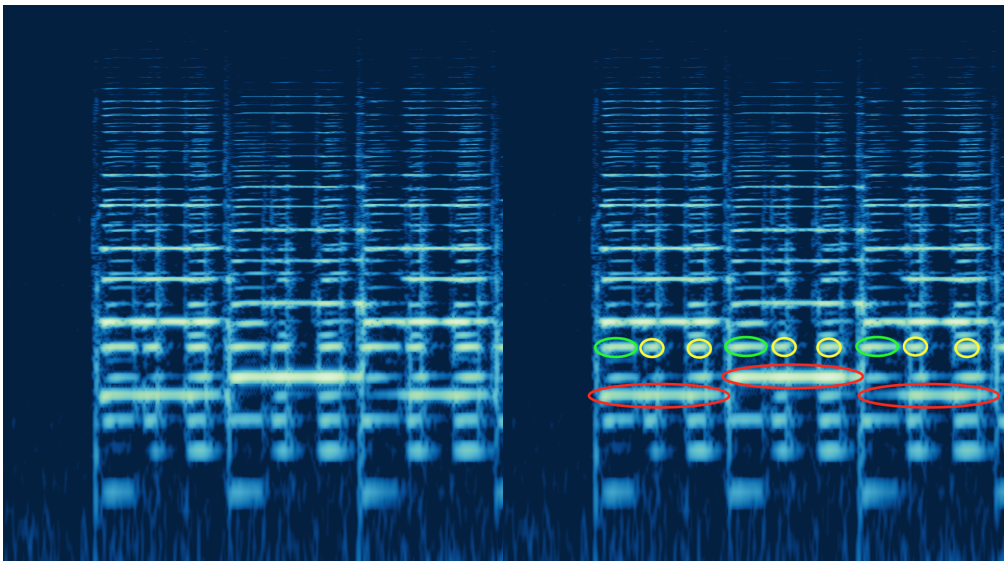
5.2 Dataset

Este trabalho propõe o FoleDataSet, contendo arquivos *wav* de interpretações das músicas de Sanfonae seus arquivos midi equivalentes, assim como o dataset

[1] O áudio utilizado nesta análise está disponível em http://bit.ly/3ttLC5g.

Figura 5.1: (a) STFT - Espectrograma - Primeira Valsa - 120bpm (b) Espectrograma com notas destacadas

(a) (b)



Fonte: Autoria Própria

MAPS e Maestro propostos por (EMIYA et al., 2010) e (HAWTHORNE et al., 2019).

Para a criação do dataset a metodologia adotada foi inspirada no trabalho de (SU; YANG, 2016), em que um músico experiente foi convidado a realizar as interpretações das músicas de forma mais fiel possível as partituras.

As partituras foram transcritas no software Musescore 3 e os arquivos midi foram exportados e ajustados em laboratório.

Devido a Sanfona tradicional ser um instrumento acústico e por isso não ter saídas eletrônicas como o piano, não foi possível utilizar uma saída midi como nos demais datasets, aumentando assim consideravelmente o trabalho para a obtenção dos dados.

O dataset também conta com gravações feitas por acordeonistas iniciantes, em que foi possível gravar a mesma música diversas vezes, realizando combinações diferentes de vozes na Sanfona, ou seja, utilizando registros diferentes e isolando a melodia do ritmo (teclado dos baixos).

Ao total o dataset conta com 33 arquivos de áudio e suas respectivas transcrições em midi, totalizando $\approx 29min$ de gravação. A Tabela 5.1 apresenta algumas informações referentes ao FoleDataset.

Música	**Compositor**	**Duração**	**Ritmo**
Primeira Valsa	Mário Mascarenhas	≈0:49s	Valsa
O Relógio Bateu 3 horas	Mário Mascarenhas	≈0:41s	Valsa
O Velhinho do Realejo	Mário Mascarenhas	≈0:51s	Valsa
Mineirinha	Mário Mascarenhas	≈0:51s	Rancheira
Parabéns para você	Patty Hill	≈0:14s	Valsa
Sanfoninha de Ouro	Mário Mascarenhas	≈0:35s	Marchinha
Uma festa no céu	Mário Mascarenhas	≈0:35s	Polka
Noite Feliz	Franz Gruber	≈0:39s	Valsa
Velhos Tempos	Desconhecido	≈0:52s	Valsa
Baião	Luiz Gonzaga / H. Teixeira	≈1:14s	Baião
Feira De Mangaio	Sivuca / G. Gadelha	≈1:26s	Baião
Olhe para o céu	Luiz Gonzaga / J. Fernandes	≈1:26s	Marchinha
Pagode Russo	Luiz Gonzaga / J. Silva	≈1:08s	Marchinha
Missioneira	Tio Bilia	≈1:18s	Vaneira
Sabiá	Luiz Gonzaga / Zé Dantas	≈1:46s	Xote

Tabela 5.1: Composição do FoleDataset v.0.0.1

Para cada música, foram gastos $\approx 90min$ para realizar a criação e alinhamento dos arquivos midi, mais $\approx 30min$ para cada áudio waveform audio format (wav) gravado pelos acordeonistas iniciantes, tempo esse principalmente devido a erros de execução durante a gravação das músicas.

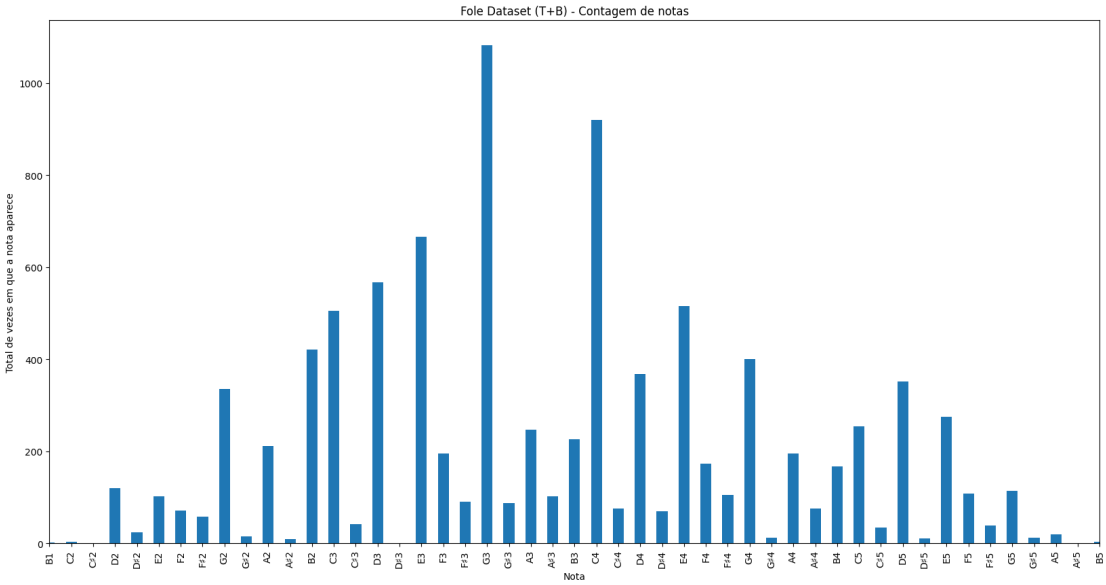
A figura 5.2 demonstra quantas vezes diferentes uma determinada nota aparece no dataset e a figura 5.3 demonstra a duração total de cada uma das notas em segundos, o dataset conta com uma média de 48s por nota com um desvio padrão de 96s e com percentis de 25%, 50%, 75% iguais a respectivamente $0s$, $2.17s$ e $43.14s$.

Cabe salientar que as figuras 5.4 e 5.5 demonstram a mesma visão das figuras 5.2 e 5.3, porém com apenas as notas do teclado, assim como as figuras 5.6 e 5.7 tem seu foco apenas nos baixos.

Observa-se com interesse que as notas do teclado tendem a se no lado direito do gráfico (A3>) enquanto os baixos tendem a se manter no lado esquerdo (A3<), com poucas notas de sobreposição. Isto é apenas uma característica do próprio dataset e das músicas que o compõe, porém, uma Sanfona de 80 baixos padrão tem 37 teclas, tendo seu início em G2, já a Sanfona de 120 baixos tem 41 teclas tendo seu início em E2, sendo assim uma heurística válida para transcrições de acordeão é que qualquer nota inferior a E2 poderia ser classificada como pertencendo aos baixos e não ao teclado, reduzindo assim a dimensionalidade

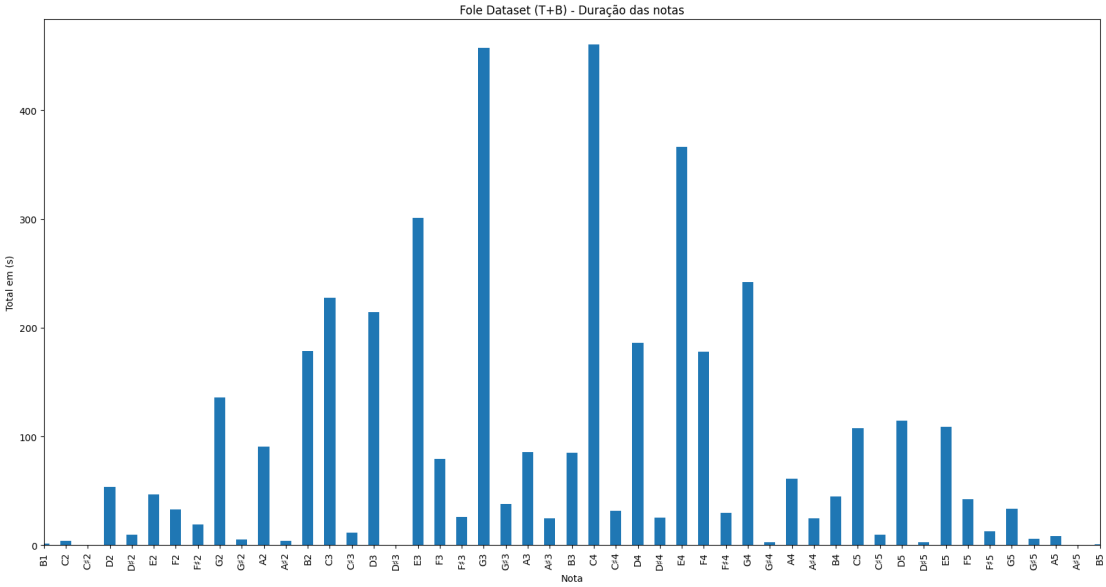
do problema.

Figura 5.2: Contagem total de cada nota no FoleDataset (Teclado+Baixos)



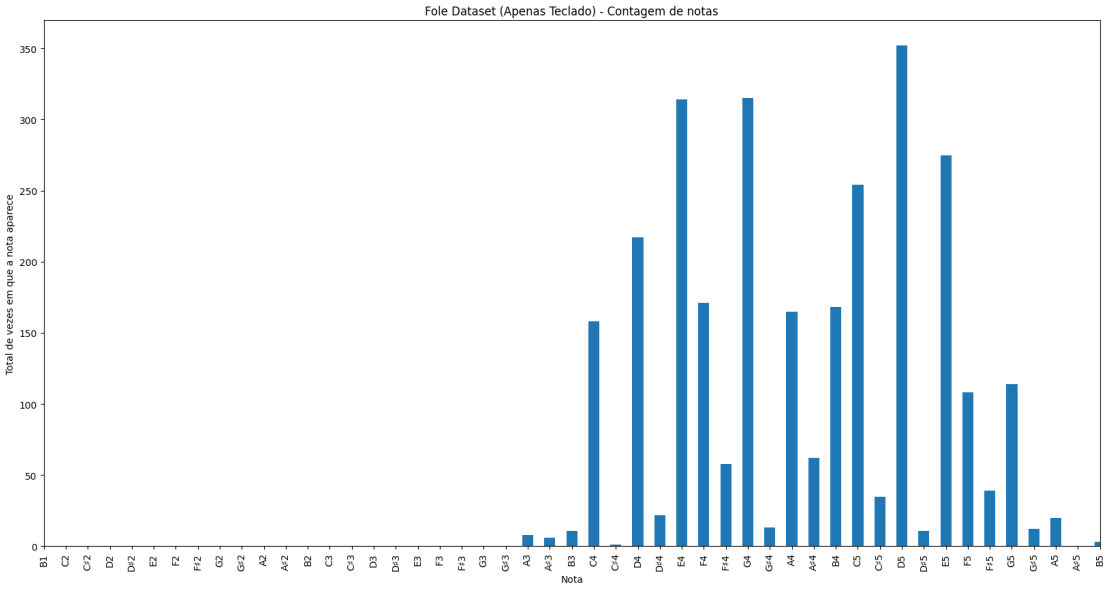
Fonte: Autoria Própria

Figura 5.3: Duração total de cada nota no FoleDataset (Teclado+Baixos)



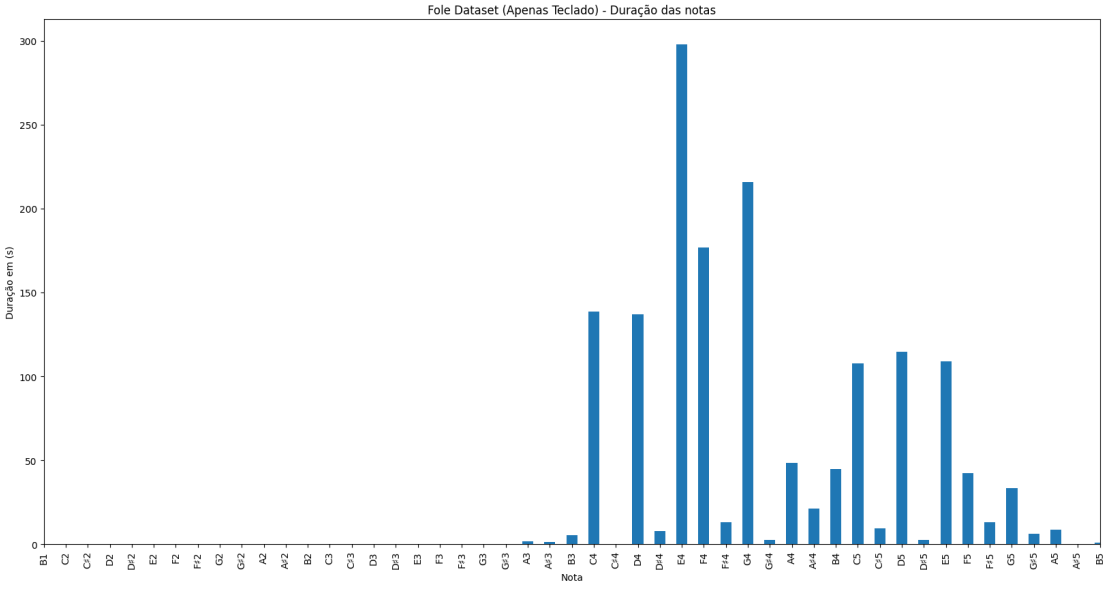
Fonte: Autoria Própria

Figura 5.4: Contagem total de cada nota no FoleDataset (Apenas Teclado)



Fonte: Autoria Própria

Figura 5.5: Duração total de cada nota no FoleDataset (Apenas Teclado)

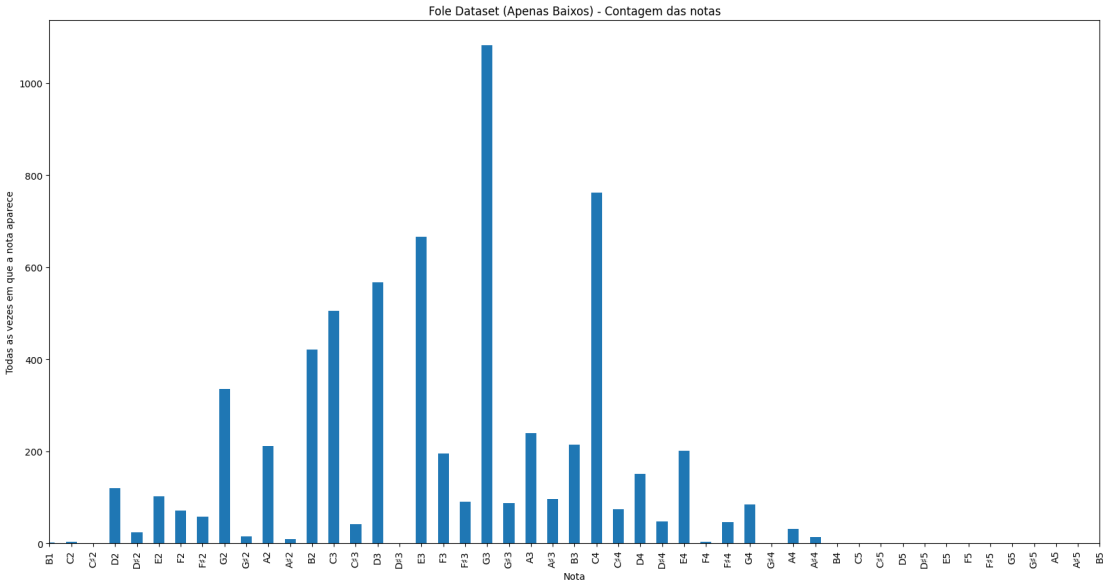


Fonte: Autoria Própria

5.3 Pré-Processamento

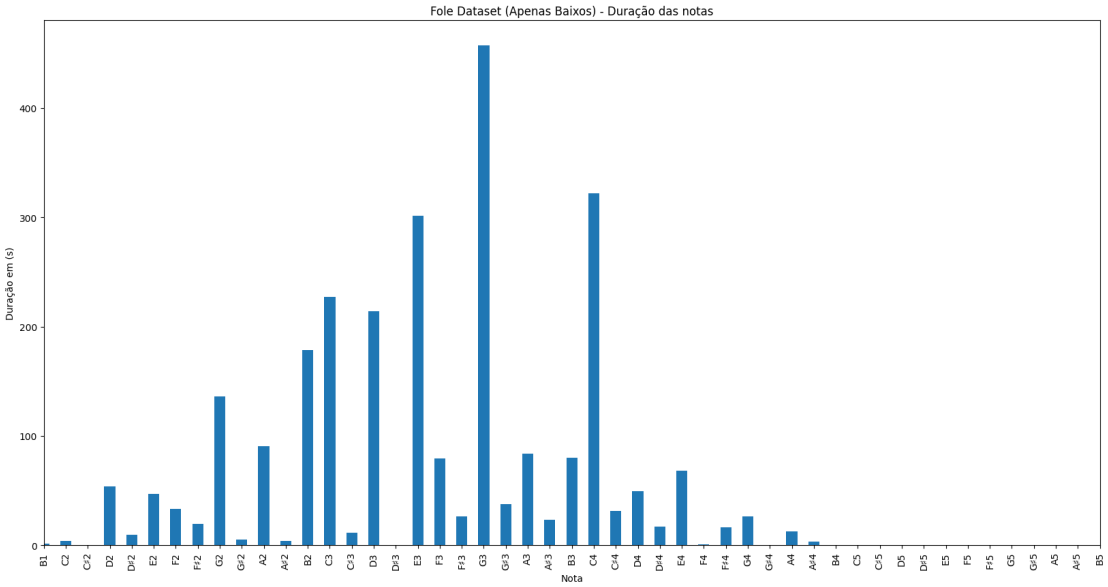
O pré-processamento para problemas de AMT é separado em duas etapas: (1) A extração de características do áudio e (2) A conversão do arquivo com as

Figura 5.6: Contagem total de cada nota no FoleDataset (Apenas Baixos)



Fonte: Autoria Própria

Figura 5.7: Duração total de cada nota no FoleDataset (Apenas Baixos)



Fonte: Autoria Própria

notações musicais, geralmente um arquivo midi, no formato de piano-roll.

5.3.1 MIDI

Como explicado na seção 2.5 o formato midi é baseado em eventos, por isso ele não pode ser processado diretamente em uma rede neural devido não se poder extrair facilmente quando uma nota inicia e termina, sendo necessário converter esse dado para um formato baseado em representação temporal, sendo o mais comum o piano-roll.

O algoritmo para a conversão é simples, o primeiro passo é definir qual a janela de tempo será utilizada. Após isso é criado um variável δ que representa quantas janelas cabem na duração da música, e então é criado um matriz preenchida com zeros de tamanho $(88, \delta)$.

Então é feita iteração por todas as mensagens do arquivo midi calculando a posição correta de início de cada nota e seu respectivo final.

Em um primeiro momento são apenas posicionados corretamente o início das notas (token 1) e o final das notas (token 3), e as colunas entre estes dois tokens são preenchidas com 2 sendo as classes que a rede deve prever, exemplificado pelo algoritmo 1 e sua saída pode ser visualizada na figura 5.8.

O formato piano-roll por se tratar de uma matriz em que cada coluna representa as notas presentes em determinada janela de tempo, desta forma pode ser utilizada mais facilmente como entrada dos algoritmos de aprendizado de máquina.

Figura 5.8: Exemplo de saída do código 1

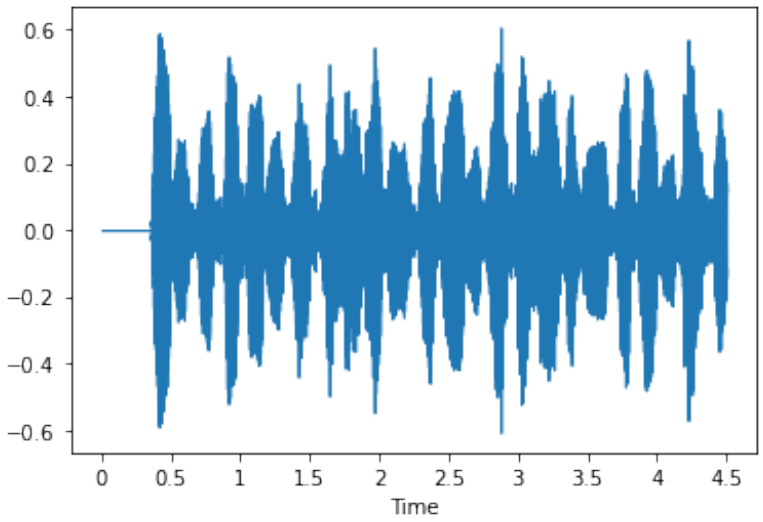
Fonte: Autoria Própria

5.3.2 Áudio

O foledataset fornece os arquivos de áudio no formato waveform audio format que é um formato de arquivo desenvolvido pela Microsoft e pela IBM como uma forma de armazenar áudio sem perca de informação *(RAW)*. O arquivo traz consigo duas informações importantes (1) o sinal de áudio e (2) taxa de gravação,

que representa quantos exemplos são necessários para completar um segundo de áudio.

Figura 5.9: Exemplo de arquivo WAV



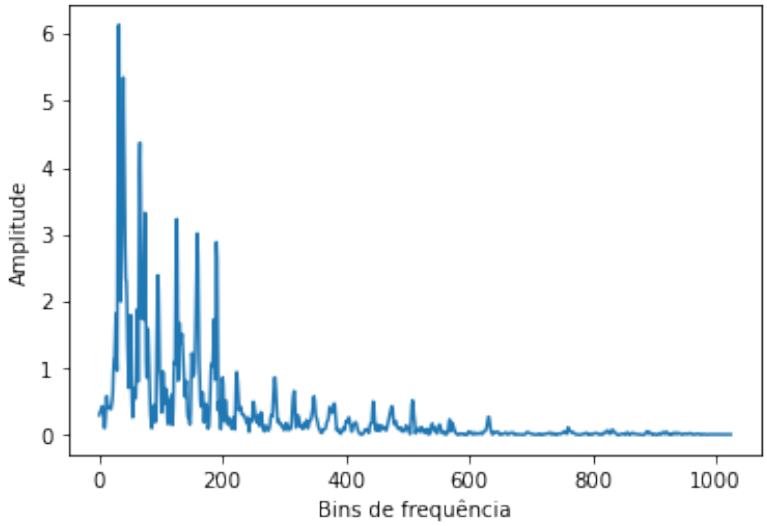
Fonte: Autoria Própria

Para o pré-processamento do áudio, é normalmente utilizado algum tipo de espectrograma para converter o problema de classificação de áudio como espectrograma de STFT também conhecido como espectrograma linear, espectrograma de Mel, Constant-Q, etc. Como este trabalho se propõe a realizar transferência de aprendizado de uma rede pré-treinada do piano, é necessário utilizar o mesmo algoritmo original do classificador, que como já foi citado na seção 4.2 é o espectrograma de Mel_{log}.

O primeiro passo para calcular este espectrograma é aplicar o filtro STFT no sinal, este filtro é uma variação do FFT com a adição de uma janela de tempo deslizante. O FFT executa um filtro no sinal convertendo o sinal baseado em tempo para um sinal baseado em frequência. A imagem 5.10 mostra uma fatia do sinal da imagem 5.9 após aplicar o filtro FFT. Infelizmente isso faz com que a propriedade tempo seja perdida, para recuperar essa propriedade é utilizado o STFT. (MONIGATTI, 2023)

Para o cálculo do STFT são utilizados dois parâmetros principais (1) n_fft que representa a largura da janela deslizante e (2) hope_length representa a quantidade amostras de áudio que será utilizada para o deslizamento da janela, a figura 5.11 demonstra como isso é aplicado em um sinal. A saída desse filtro é

Figura 5.10: Exemplo de espectrograma FFT



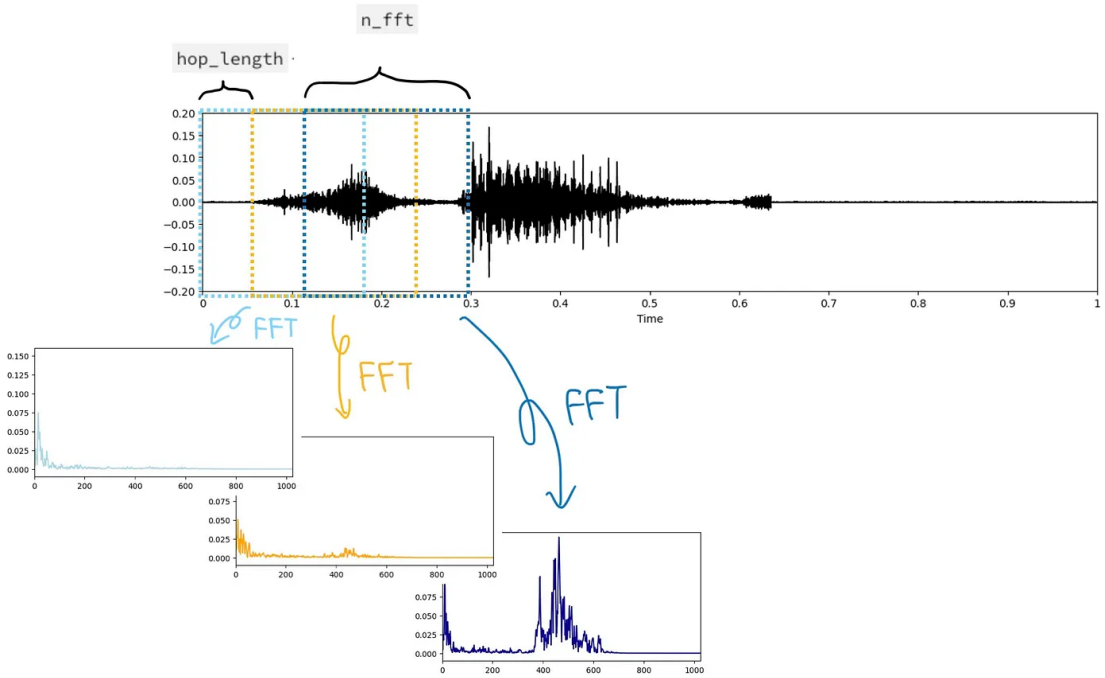
Fonte: Autoria Própria

o espectrograma de STFT também conhecido como espectrograma linear. Neste espectrograma, por sua vez, é aplicado um filtro que converte a amplitude para a escala de mel explicada anteriormente na seção 2. (MONIGATTI, 2023)

O parâmetro hope_length também define a largura da matriz, ou imagem, resultante do espectrograma, sendo a largura definida por $w = \lceil \frac{(sr \times audio\_length)}{hop\_length} \rceil$, onde sr é a taxa de gravação do áudio e $audio\_length$ a largura do sinal de áudio. A altura da matriz, ou imagem, é definido pelo número de mels presentes no espectrograma, na rede proposta pelo grupo magenta explicada na seção 4.2 ao analisar o código foi constatado que o espectrograma de mel é calculado com o uso de 229 mels, sendo assim possível inferir que o formato de entrada desta rede é uma matriz de $(229, 1)$.

Na rede onset-and-frames do grupo magenta é utilizando um $sr = 16000$ e um hop_length de 512, ao fazer o cálculo reverso $d = \frac{16000}{512}$ é possível constatar que cada segundo do áudio representa $31,25$ colunas, sendo assim $\frac{1}{31,25} = 0,032s$. É importante ressaltar que se deve utilizar a mesma escala de tempo na criação do piano-roll explicada na seção 2.5.

Figura 5.11: Cálculo do STFT



Fonte: (MONIGATTI, 2023)

5.4 Validação

O método de validação utilizado neste trabalho, devido à baixa quantidade de dados disponíveis de músicas de Sanfona, será o leave-one-subject-out-cross-validation. Neste método de validação, são realizados N treinamentos, em cada treinamento uma instância diferente é utilizada na base de teste. Diferente do leave-one-out-cross-validation as instâncias semelhantes da base de dados, músicas que possuem mais de uma variação, são consideradas uma única instância na separação da base. (GHOLAMIANGONABADI; KISELOV; GROLINGER, 2020)

5.5 Avaliação

O método mais utilizado para avaliar sistemas de AMT no estado da arte consiste na técnica proposta por (BAY; EHMANN; DOWNIE, 2009), denominada como F1-Score. Este método considera notas detectadas corretamente se o

semi-tom, inicio e duração da nota estiverem corretos com uma margem de erro de no máximo $50ms$. Estas notas são consideradas VP (verdadeiros positivos), quaisquer outras notas detectadas são FP (falsos positivos) e notas não detectadas são consideradas FN (falsos negativos) e é calculado o $F1 - Score$ destas informações.

$$Precision = \frac{\sum_{t=1}^{T} VP_t}{\sum_{t=1}^{T} VP_t + FP_t} \tag{5.1}$$

$$Recall = \frac{\sum_{t=1}^{T} VP_t}{\sum_{t=1}^{T} VP_t + FN_t} \tag{5.2}$$

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.3}$$

Esta métrica, no entanto, não considera diversas informações que podem impactar em uma transcrição completa. Assim como citado por (HAWTHORNE et al., 2018), devido a isso foi incluída outra métrica para avaliação do modelo baseado no trabalho feito por (MCLEOD; STEEDMAN, 2018) que introduziu uma métrica chamada $MV2H$. [2]

A métrica $MV2H$ é uma média aritmética de outras métricas, sendo elas: *Multi-Pitch Detection (f1-score), separação da voz, Metrical alignment, Note Value Detection e Harmonic Analysis*.

Apesar da métrica $MV2H$ ser uma métrica mais completa, como este trabalho realiza a transferência de aprendizado utilizando a rede proposta pelo grupo magenta, explicada na seção 4.2 é utilizado apenas a f0 como proposta originalmente no trabalho do magenta.

[2]Para este trabalho a métrica foi re-implementada em Python e disponibilizada em https://github.com/lucasmpaim/pyMV2H.

6

Experimentos

Neste capítulo são apresentados os experimentos realizados durante a execução deste trabalho. Em cada experimento este trabalho irá tentar responder as seguintes questões de pesquisa:

- É possível realizar a transferência de aprendizado com um classificador originalmente treinado para o piano para um novo classificador para Sanfona?
- Congelar diferentes números de camadas convolucionais altera a qualidade da transcrição? E qual o impacto no tempo de treinamento?
- Qual o impacto de se treinar a rede por mais épocas?
- Os diferentes registros da Sanfona impactam diretamente na transcrição?
- Qual o impacto do treinamento utilizando a base de Sanfona no problema original?
- Os baixos (botões do lado esquerdo da Sanfona) impactam no desempenho do modelo?

6.1 Experimento #01: Baseline

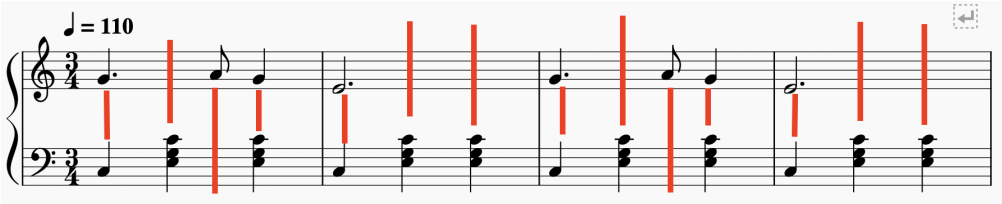
O objetivo deste experimento é definir métricas a serem utilizadas como base de comparação dos demais experimentos tanto para o FoleDataset proposto por este trabalho quanto para o Maestro Dataset que é o dataset original do modelo utilizado.

Neste experimento foi utilizado apenas o modelo pré-treinado do magenta, explicado em detalhes na seção 4.2, sem nenhum treinamento extra.

A tabela 6.1 apresenta o f1-score, precision e recall do fole dataset. A música com a pior transcrição geral (considerando o f1-score), é a valsa "Velhos tempos" com apenas 0, 6% de acerto e a música que teve a melhor transcrição é outra valsa "Noite Feliz" atingindo 34, 9% de acerto.

Apesar de ambas as músicas terem o mesmo ritmo base, a valsa, as músicas seguem uma estrutura diferente de compasso que torna a transcrição da música "Noite Feliz" seja uma tarefa mais simples. A figura 6.1 exemplifica a estrutura seguida nesta música onde muitas notas (teclado e baixos) são tocadas isoladamente durante a execução da música.

Figura 6.1: Estrutura Noite Feliz

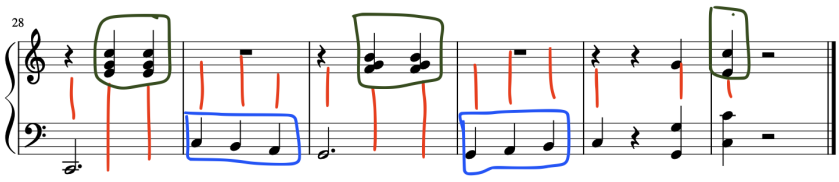


Fonte: Autoria Própria

Em contrapartida, a música "Velhos tempos" possui uma estrutura muito mais complexa e particularidades em relação ao restante do dataset, como:

- Formação de acordes no teclado, apesar da música "Missioneiro" também apresentar acordes, na música "Velhos tempos" eles estão muito mais presentes.
- Escala nos baixos, a música é a única no dataset que apresenta a execução de escalas nos baixos.

Figura 6.2: Estrutura velhos tempos



Fonte: Autoria Própria

A figura 6.2 mostra um trecho da valsa "Velhos Tempos" com alguns trechos ressaltados, os círculos em verde-escuro mostram acordes na mão direita (teclado) e os círculos em azul mostram escalas sendo tocadas na mão esquerda (baixos).

Música	**f1-score**	**precision**	**recall**
primeira valsa	0.17715	0.12520	0.30278
baião	0.27313	0.19826	0.43884
feira de mangaio	0.21364	0.21057	0.21680
mineirinha	0.14088	0.12068	0.16918
missioneiro	0.14103	0.12203	0.16704
noite feliz	0.34944	0.29936	0.41964
o velhinho do realejo	0.19444	0.15625	0.25735
o relógio bateu 3 horas	0.17496	0.12745	0.27896
olha para o céu	0.29432	0.25922	0.34042
pagode russo	0.12181	0.11178	0.13381
parabéns	0.26136	0.23711	0.29113
sabiá	0.31879	0.24520	0.45548
sanfoninha de ouro	0.24327	0.20077	0.30860
uma festa no céu	0.20210	0.16637	0.25737
velhos tempos	0.00698	0.00583	0.00869

Tabela 6.1: Métricas experimento #01 - Baseline

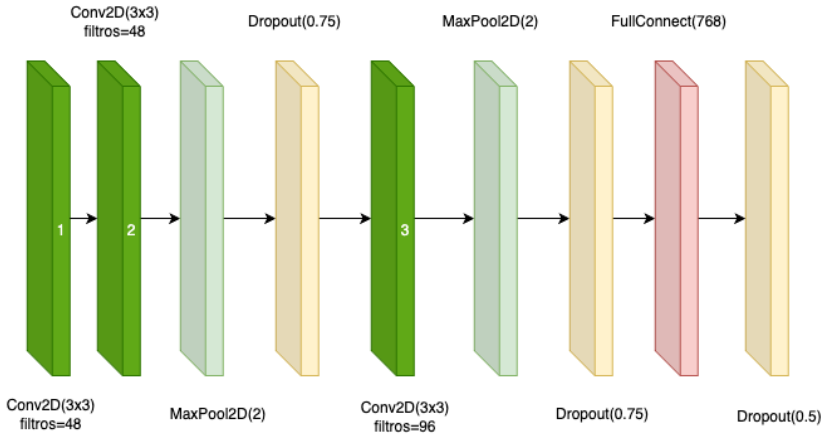
6.2 Experimento #02: Fine Tunning

O objetivo deste experimento foi verificar o desempenho das transcrições utilizando uma rede pré-treinada para o piano, realizando apenas novas iterações de treinamento utilizando a base proposta por este trabalho, a FoleDataSet.

A figura 6.3 demonstra a configuração da pilha convolucional, neste experimento não foi congelada nenhuma camada.

A tabela 6.2 demonstra os resultados após o treinamento de 3300, 6600 e 9900 épocas, nela é possível notar que houve uma melhora no precision na maioria das músicas, mostrando que quando uma nota é detectada pelo modelo é maior a chance do modelo estar correto, ou seja, uma diminuição significativa nos falsos positivos do modelo, porém, com a diminuição do recall é possível notar um aumento nos falsos negativos, ou seja, a rede tende a ignorar mais as notas tocadas no acordeon. Também é possível notar que houve uma leve melhora do recall neste experimento quando o treinamento atingiu 9900 épocas.

Figura 6.3: Pilha Convolucional



Fonte: Autoria Própria

Com 3300 épocas houve uma melhora em f1 de 8 músicas e uma piora em 7 músicas, precision teve uma melhora em 13 músicas e uma piora em duas e recall apresentou uma melhora em 4 músicas e uma piora em 11, apresentando um ganho médio de ≈ 2% em f1, um ganho médio de ≈ 11% em precision e uma perca média de ≈ 7% em recall.

Com 6600 épocas houve uma melhora em f1 de 8 músicas e uma piora em 7 músicas, precision teve uma melhora em 14 músicas e uma piora em uma e recall apresentou uma melhora em 4 músicas e uma piora em 11, apresentando um ganho médio de ≈ 2,8% em f1, um ganho médio de ≈ 20% em precision e uma perca média de ≈ 8% em recall.

Com 9900 épocas houve uma melhora em f1 de 8 músicas e uma piora em 7 músicas, precision teve uma melhora em 14 músicas e uma piora em uma e recall apresentou uma melhora em 3 músicas e uma piora em 12, apresentando um ganho médio de ≈ 4% em f1, um ganho médio de ≈ 24% em precision e uma perca média de ≈ 8% em recall.

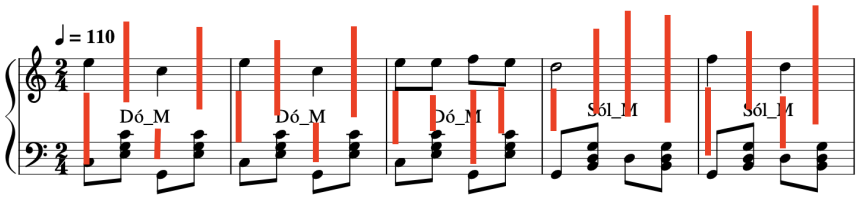
A tabela 6.3 mostra a variação das métricas deste experimento em relação ao experimento #01 - Baseline.

A música que obteve a melhor transcrição neste experimento foi a marchinha Sanfoninha de ouro que obteve uma melhora de ≈ 29% em f1, de ≈ 53% em precision e de ≈ 11% em recall.

A música baião teve uma piora de ≈ 15% em f1, uma melhora de ≈ 10% em precision e uma piora ≈ 36% em recall.

A marchinha "sanfoninha de ouro" por ser uma música de introdução a Sanfona tem uma estrutura simples se compararmos com outras músicas do dataset, como demonstrado pela figura 6.4.

Figura 6.4: Estrutura Sanfoninha de ouro

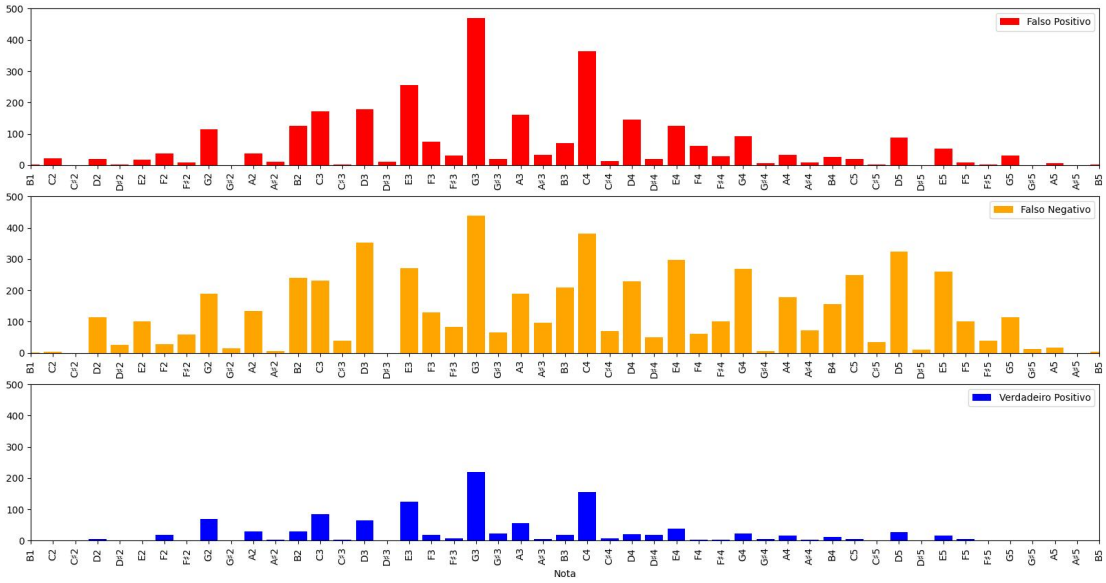


Fonte: Autoria Própria

As figuras 6.5, 6.6 e 6.7 demonstram o histograma de notas do experimento, separados em: Falso positivo (vermelho), falso negativo (laranja) e verdadeiro positivo (azul), a partir destas imagens é possível notar que o modelo tende a errar mais as notas G3, C4, porém curiosamente, são as notas que o modelo tende a acertar mais.

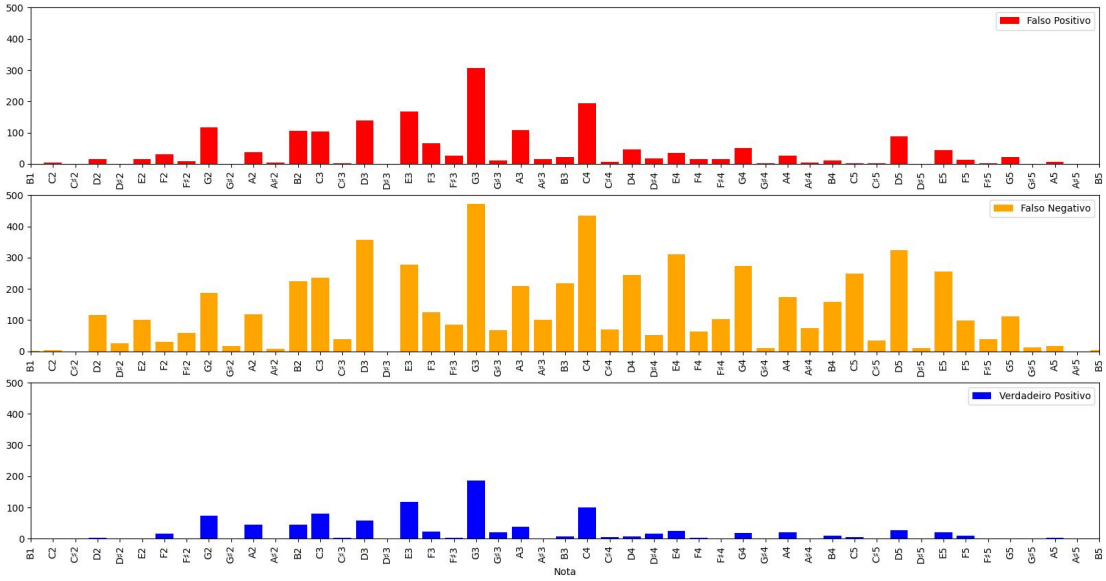
Conforme a quantidade de épocas aumenta é possível notar que os falsos positivos (barras vermelhas) tendem a diminuir (aumentando o precision) enquanto os falsos negativos tendem a aumentar (diminuindo o recall) com exceção das 9900 épocas que mostrou um aumento de recall.

Figura 6.5: Histograma - Experimento #02 - 3300



Fonte: Autoria Própria

Figura 6.6: Histograma - Experimento #02 - 6600



Fonte: Autoria Própria

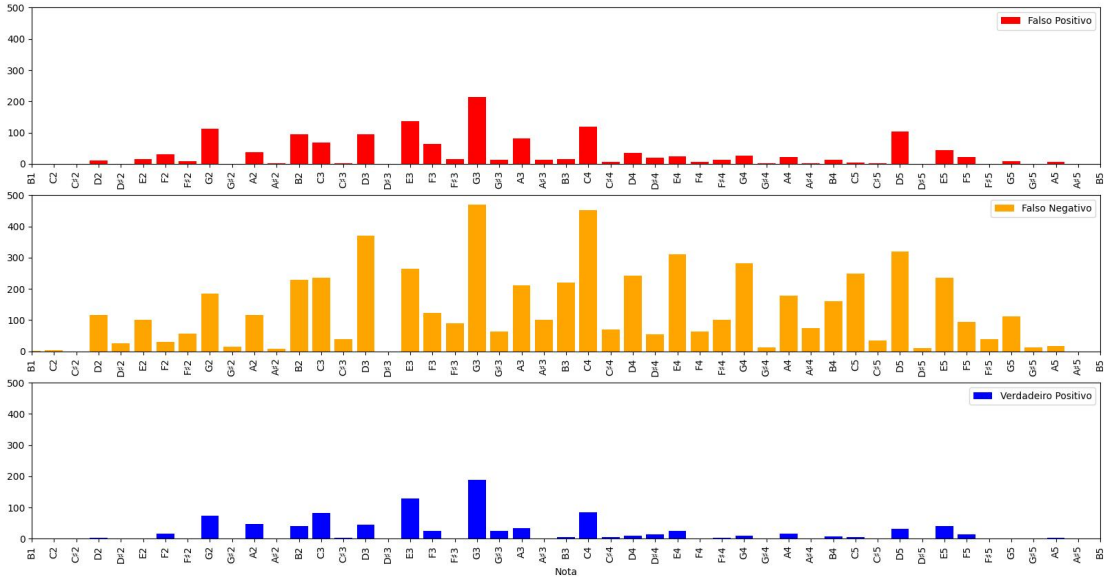
	3300 Épocas			6600 Épocas			9900 Épocas		
Música	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
Primeira Valsa	0.321370	0.281440	**0.374500**	**0.349080**	0.360170	0.338650	0.313730	**0.407640**	0.254980
Baião	**0.172960**	0.251140	**0.131890**	0.159570	**0.306120**	0.107910	0.119230	0.300970	0.074340
Feira de Mangaio	**0.210010**	0.386190	**0.144220**	0.201040	0.456080	0.128940	0.208710	**0.487720**	0.132760
Mineirinha	0.189370	0.210330	0.172210	0.223830	0.278030	0.187310	**0.247190**	**0.325120**	**0.199400**
Missioneiro	**0.073310**	0.126720	**0.051570**	0.047880	0.134020	0.029150	0.051330	**0.140700**	0.031390
Noite Feliz	0.353260	0.451390	0.290180	0.385060	0.540320	**0.299110**	**0.391440**	**0.621360**	0.285710
Velinho do realejo	0.284460	0.351350	**0.238970**	**0.324740**	0.543100	0.231620	0.320440	**0.644440**	0.213240
Relógio bateu 3 horas	0.247090	0.270410	0.227470	0.256410	0.381360	0.193130	**0.307260**	**0.440000**	**0.236050**
Olha pro céu	**0.221260**	0.411580	**0.151300**	0.162520	**0.425000**	0.100470	0.150197	0.457831	0.08983
Pagode russo	**0.103560**	0.154590	**0.077860**	0.080790	0.154110	0.054740	0.066150	**0.165050**	0.041360
Parabéns	0.292310	0.372550	0.240510	0.322030	0.487180	0.240510	**0.382610**	**0.611110**	**0.278480**
Sabiá	**0.173000**	0.240740	**0.135010**	0.143920	0.250920	0.100890	0.155760	**0.325470**	0.102370
Sanfoninha de ouro	0.397280	0.464290	0.347180	0.476890	0.632350	0.382790	**0.533080**	**0.734380**	**0.418400**
Uma festa no céu	0.374470	0.397590	0.353890	0.406090	0.550460	0.321720	**0.490630**	**0.672900**	**0.386060**
Velhos tempos	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**

Tabela 6.2: Resultados Fine Tunning

	3300 Épocas			6600 Épocas			9900 Épocas		
Música	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
baião	-0.10017	+0.05287	-0.30696	-0.11356	+0.10785	-0.33094	-0.15390	+0.10270	-0.36451
feira de mangaio	-0.00364	+0.17561	-0.07259	-0.01261	+0.24550	-0.08787	-0.00494	+0.27714	-0.08405
primeira valsa	+0.14421	+0.15623	+0.07171	+0.17192	+0.23496	+0.03586	+0.13657	+0.28243	-0.04781
mineirinha	+0.04849	+0.08964	+0.00303	+0.08295	+0.15734	+0.01813	+0.10631	+0.20443	+0.03022
missioneiro	-0.06772	+0.00469	-0.11547	-0.09315	+0.01199	-0.13789	-0.08970	+0.01867	-0.13565
noite feliz	+0.00382	+0.15203	-0.12946	+0.03562	+0.24096	-0.12053	+0.04200	+0.32200	-0.13393
o relógio bateu 3 horas	+0.07212	+0.14296	-0.05150	+0.08144	+0.25391	-0.08584	+0.13229	+0.31255	-0.04292
o velhinho do realejo	+0.09002	+0.19510	-0.01838	+0.13030	+0.38685	-0.02573	+0.12600	+0.48819	-0.04411
olha para o céu	-0.07307	+0.15235	-0.18913	-0.13181	+0.16577	-0.23996	-0.14413	+0.19861	-0.25059
pagode russo	-0.01826	+0.04280	-0.05596	-0.04103	+0.04232	-0.07908	-0.05567	+0.05326	-0.09246
parabéns	+0.03095	+0.13544	-0.05063	+0.06067	+0.25007	-0.05063	+0.12125	+0.37400	-0.01266
sabia	-0.14580	-0.00447	-0.32048	-0.17488	+0.00571	-0.35460	-0.16304	+0.08026	-0.35312
sanfoninha de ouro	+0.15401	+0.26352	+0.03857	+0.23362	+0.43158	+0.07418	+0.28981	+0.53361	+0.10979
uma festa no céu	+0.17236	+0.23121	+0.09652	+0.20398	+0.38408	+0.06435	+0.28852	+0.50652	+0.12869
velhos tempos	-0.00698	-0.00583	-0.00870	-0.00698	-0.00583	-0.00870	-0.00698	-0.00583	-0.00870

Tabela 6.3: Resultados Fine Tunning - Comparação com Baseline

Figura 6.7: Histograma - Experimento #02 - 9900



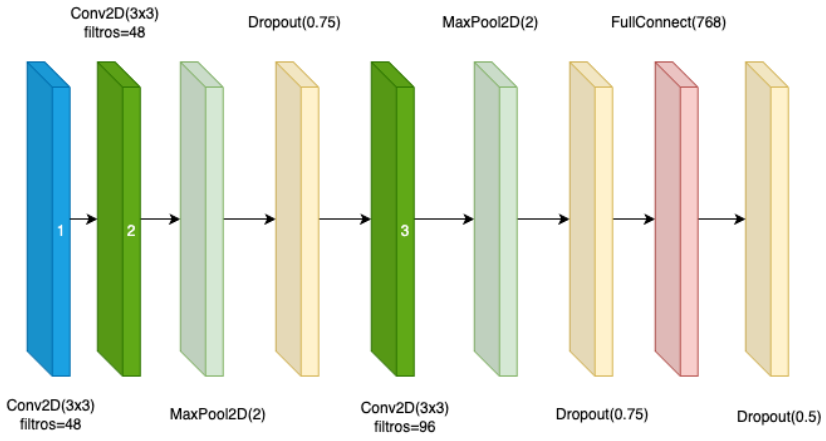
Fonte: Autoria Própria

6.3 Experimento #03: Transferência de aprendizado - 1 Camada congelada

O objetivo deste experimento é demonstrar as métricas aplicando a técnica de transferência de aprendizado de transferência por representação de características como explicado na seção 3.2.2.

Neste experimento foi congelada a primeira camada convolucional da pilha convolucional do modelo como demonstrado na figura 6.8 e realizado o treinamento por 3300, 6600 e 9900 épocas.

Figura 6.8: Pilha Convolucional - 1 Camada congelada



Fonte: Autoria Própria

A tabela 6.4 demonstra que assim como o experimento #02 este experimento também demonstra uma diminuição consistente dos falsos positivos (precision) e um aumento consistente dos falsos negativos (recall).

Assim como no experimento #02 a música com o melhor desempenho foi a marchinha "sanfoninha de ouro" e a de pior desempenho foi a valsa "velhos tempos".

Com 3300 épocas houve uma melhora em f1 de 7 músicas e uma piora em 8 músicas, precision teve uma melhora em 14 músicas e uma piora em apenas uma e recall apresentou uma melhora em 3 músicas e uma piora em 10, apresentando um ganho médio de $\approx 1\%$ em f1, um ganho médio de $\approx 11\%$ em precision e uma perca média de $\approx 8\%$ em recall.

Com 6600 épocas houve uma melhora em f1 de 8 músicas e uma piora em 7 músicas, precision teve uma melhora em 13 músicas e uma piora em duas e

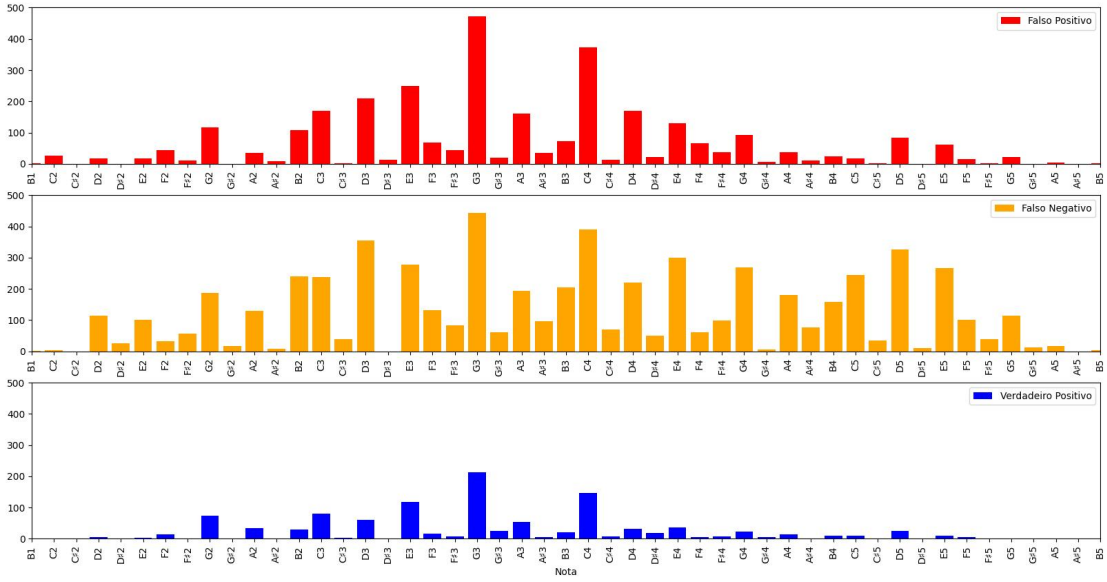
recall apresentou uma melhora em 3 músicas e uma piora em 12, apresentando um ganho médio de ≈ 2,4% em f1, um ganho médio de ≈ 17% em precision e uma perca média de ≈ 8% em recall.

Com 9900 épocas houve uma melhora em f1 de 7 músicas e uma piora em 8 músicas, precision teve uma melhora em 14 músicas e uma piora em uma e recall apresentou uma melhora em 3 músicas e uma piora em 12, apresentando um ganho médio de ≈ 2,7% em f1, um ganho médio de ≈ 22% em precision e uma perca média de ≈ 9,6% em recall.

A tabela 6.5 demonstra a diferença por música em comparação ao experimento baseline.

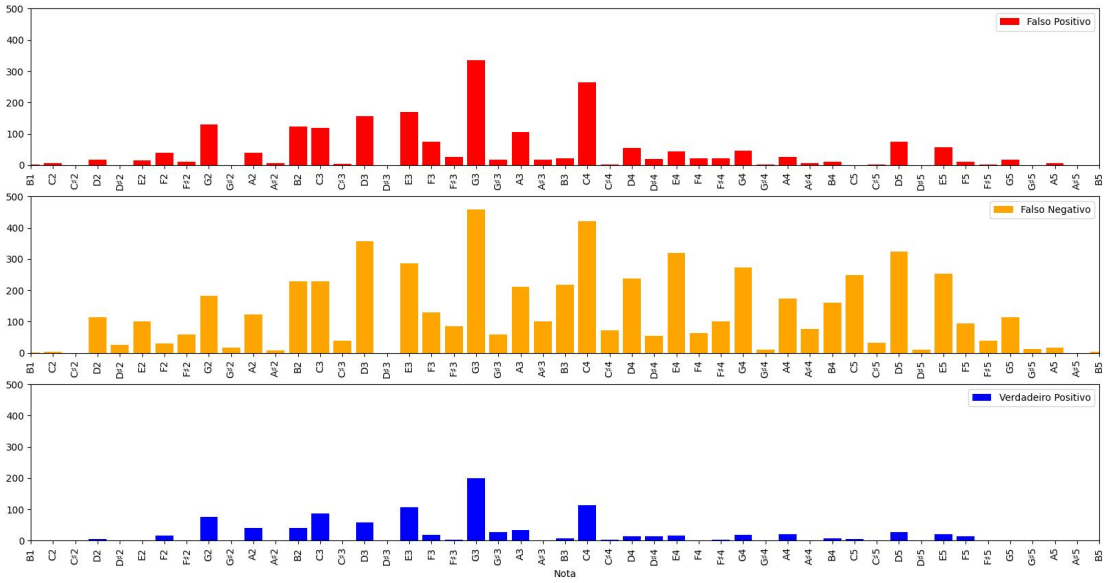
As figuras 6.9, 6.10 e 6.11 apresentam os histogramas deste experimento, que assim como o experimento anterior, demonstram a mesma tendência de diminuição dos falsos positivos e aumento dos falsos negativos conforme o treinamento avança, porém diferente do experimento anterior, ao chegar em 9900 épocas os falsos negativos continuaram aumentando.

Figura 6.9: Histograma - Experimento #03 - 3300



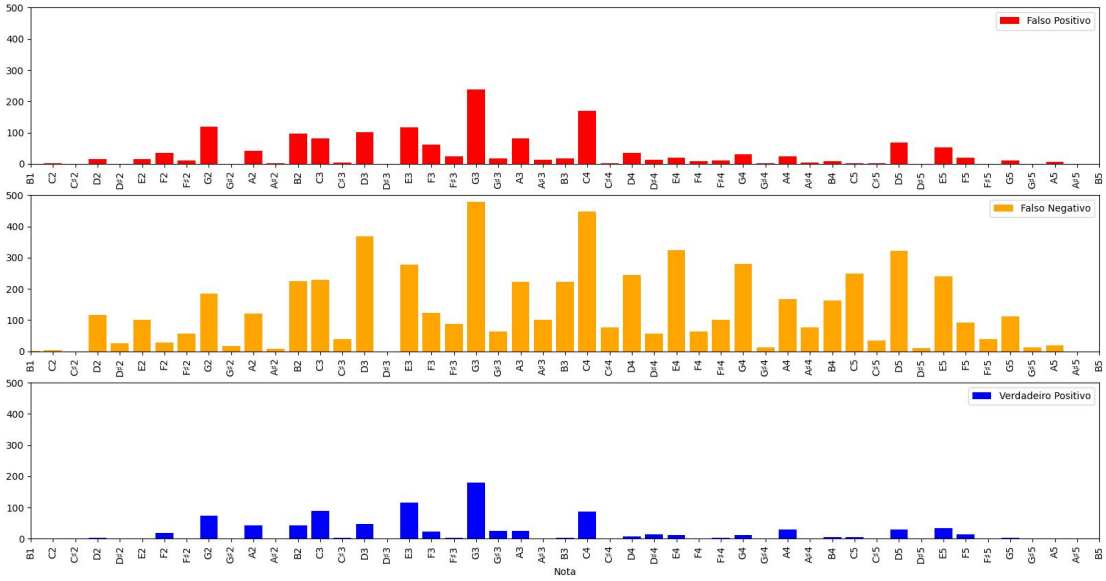
Fonte: Autoria Própria

Figura 6.10: Histograma - Experimento #03 - 6600



Fonte: Autoria Própria

Figura 6.11: Histograma - Experimento #03 - 9900



Fonte: Autoria Própria

	3300 Épocas			6600 Épocas			9900 Épocas		
Música	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
Primeira valsa	0.310470	0.283830	0.342630	**0.368420**	0.374490	**0.362550**	0.317070	**0.408810**	0.258960
Baião	**0.160120**	0.216330	**0.127100**	0.140940	0.234640	0.100720	0.129630	**0.284550**	0.083930
feira de mangaio	0.211980	0.379310	**0.147090**	**0.212670**	0.447850	0.139450	0.202830	**0.462590**	0.129890
mineirinha	0.162990	0.186050	0.145020	0.201470	0.255810	0.166160	**0.243710**	**0.338710**	**0.190330**
missioneiro	**0.083400**	**0.124170**	**0.062780**	0.054560	0.113880	0.035870	0.042590	0.122340	0.025780
noite feliz	0.343010	0.419350	**0.290180**	**0.353280**	0.488190	0.276790	0.347310	**0.527270**	0.258930
o velhinho do realejo	0.266380	0.327960	0.224260	0.295400	0.432620	0.224260	**0.326630**	**0.515870**	**0.238970**
o relógio bateu 3 horas	0.262670	0.283580	**0.244640**	0.264370	0.400000	0.197420	**0.267060**	**0.432690**	0.193130
olha para o céu	**0.199820**	0.407270	**0.132390**	0.167150	**0.446150**	0.102840	0.125130	0.427590	0.073290
pagode russo	**0.095930**	0.144610	**0.071780**	0.074410	0.146430	0.049880	0.062440	**0.157640**	0.038930
parabens	0.303030	0.377360	**0.253160**	**0.324790**	0.500000	0.240510	0.317760	**0.607140**	0.215190
sabia	**0.183560**	0.258060	**0.142430**	0.150700	0.274510	0.103860	0.145290	**0.309180**	0.094960
sanfoninha de ouro	0.397350	0.449440	0.356080	0.472530	0.617220	0.382790	**0.531840**	**0.720810**	**0.421360**
uma festa no céu	0.352590	0.394040	0.319030	0.398740	0.481060	0.340480	**0.461030**	**0.604350**	**0.372650**
velhos tempos	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**

Tabela 6.4: Resultados do experimento #03 - 1 Camada Congelada

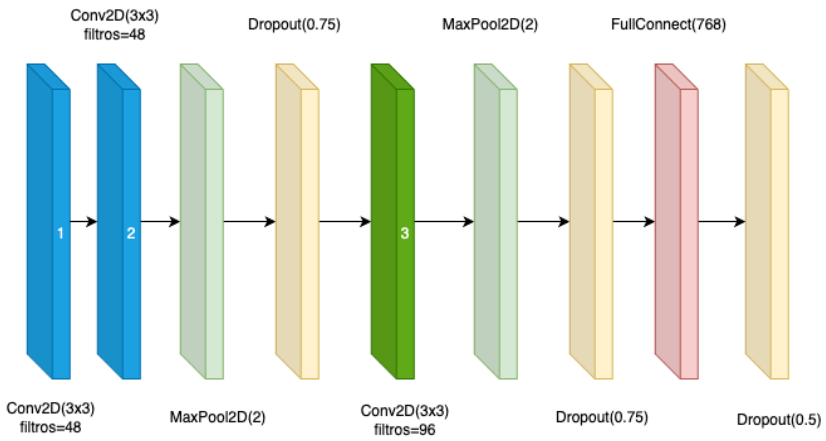
	3300 Épocas			6600 Épocas			9900 Épocas		
Música	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
baião	-0.11301	+0.01806	-0.31175	-0.13219	+0.03637	-0.33813	-0.14350	+0.08628	-0.35492
feira de mangaio	-0.00167	+0.16873	-0.06972	-0.00098	+0.23727	-0.07736	-0.01082	+0.25201	-0.08692
primeira valsa	+0.13331	+0.15862	+0.03984	+0.19126	+0.24928	+0.05976	+0.13991	+0.28360	-0.04383
mineirinha	+0.02211	+0.06536	-0.02416	+0.06059	+0.13512	-0.00302	+0.10283	+0.21802	+0.02115
missioneiro	-0.05763	+0.00214	-0.10426	-0.08647	-0.00815	-0.13117	-0.09844	+0.00031	-0.14126
noite feliz	-0.00643	+0.11999	-0.12946	+0.00384	+0.18883	-0.14285	-0.00213	+0.22791	-0.16071
o relógio bateu 3 horas	+0.08770	+0.15613	-0.03433	+0.08940	+0.27255	-0.08155	+0.09209	+0.30524	-0.08584
o velhinho do realejo	+0.07194	+0.17171	-0.03309	+0.10096	+0.27637	-0.03309	+0.13219	+0.35962	-0.01838
olha para o céu	-0.09451	+0.14804	-0.20804	-0.12718	+0.18692	-0.23759	-0.16920	+0.16836	-0.26714
pagode russo	-0.02589	+0.03282	-0.06204	-0.04741	+0.03464	-0.08394	-0.05938	+0.04585	-0.09489
parabéns	+0.04167	+0.14025	-0.03798	+0.06343	+0.26289	-0.05063	+0.05640	+0.37003	-0.07595
sabia	-0.13524	+0.01285	-0.31306	-0.16810	+0.02930	-0.35163	-0.17351	+0.06397	-0.36053
sanfoninha de ouro	+0.15408	+0.24867	+0.04747	+0.22926	+0.41645	+0.07418	+0.28857	+0.52004	+0.11275
uma festa no ceu	+0.15048	+0.22766	+0.06166	+0.19663	+0.31468	+0.08311	+0.25892	+0.43797	+0.11528
velhos tempos	-0.00698	-0.00583	-0.00870	-0.00698	-0.00583	-0.00870	-0.00698	-0.00583	-0.00870

Tabela 6.5: Resultados do experimento #03 - Comparação com Baseline

6.4 Experimento #04: 2 Camadas congeladas

O objetivo deste experimento é avaliar o modelo aplicando a técnica de transferência de aprendizado chamada transferência de representação de características como explicado na seção 3.2.2. Neste experimento foram congeladas duas camadas convolucionais como demonstrado pela figura 6.12 e realizado o treinamento por 3300, 6600 e 9900 épocas.

Figura 6.12: Pilha Convolucional - 2 Camadas convolucionais congeladas



Fonte: Autoria Própria

A tabela 6.6 demonstra os resultados deste experimento, é possível notar uma melhora consistente na precisão do modelo, ou seja, comparado aos outros experimentos houve uma diminuição dos falsos positivos.

Com exceção das músicas “Mineirinha”, “Sanfoninha de ouro” e “Uma festa no céu” houve uma redução do recall, ou seja, houve um aumento nos falsos negativos, ou seja, a rede tende a ignorar mais notas que estão sendo tocadas no instrumento.

Com 3300 épocas houve uma melhora em f1 de 7 músicas e uma piora em 8 músicas, precision teve uma melhora em 12 músicas e uma piora em 3 e recall apresentou uma melhora em 5 músicas e uma piora em 10, apresentando um ganho médio de ≈ 1% em f1, um ganho médio de ≈ 10% em precision e uma perca média de ≈ 7% em recall.

Com 6600 épocas houve uma melhora em f1 de 8 músicas e uma piora em 7 músicas, precision teve uma melhora em 13 músicas e uma piora em duas e recall apresentou uma melhora em 4 músicas e uma piora em 11, apresentando

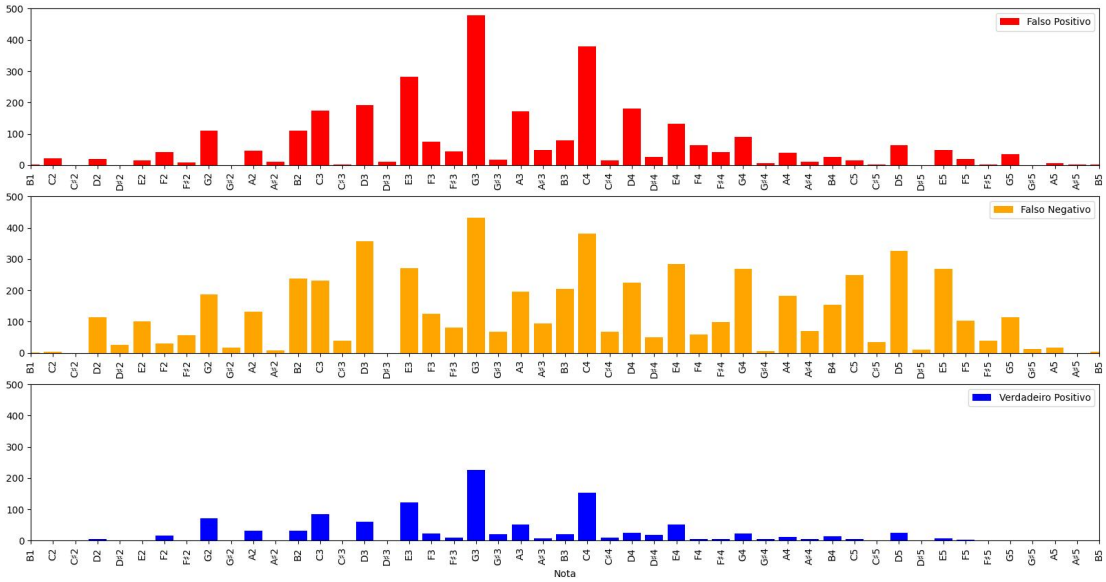
um ganho médio de ≈ 2,8% em f1, um ganho médio de ≈ 17% em precision e uma perca média de ≈ 8% em recall.

Com 9900 épocas houve uma melhora em f1 de 7 músicas e uma piora em 8 músicas, precision teve uma melhora em 13 músicas e uma piora em duas e recall apresentou uma melhora em 3 músicas e uma piora em 12, apresentando um ganho médio de ≈ 2,7% em f1, um ganho médio de ≈ 21% em precision e uma perca média de ≈ 9,5% em recall.

A tabela 6.7 demonstra a diferença por música em comparação ao experimento baseline.

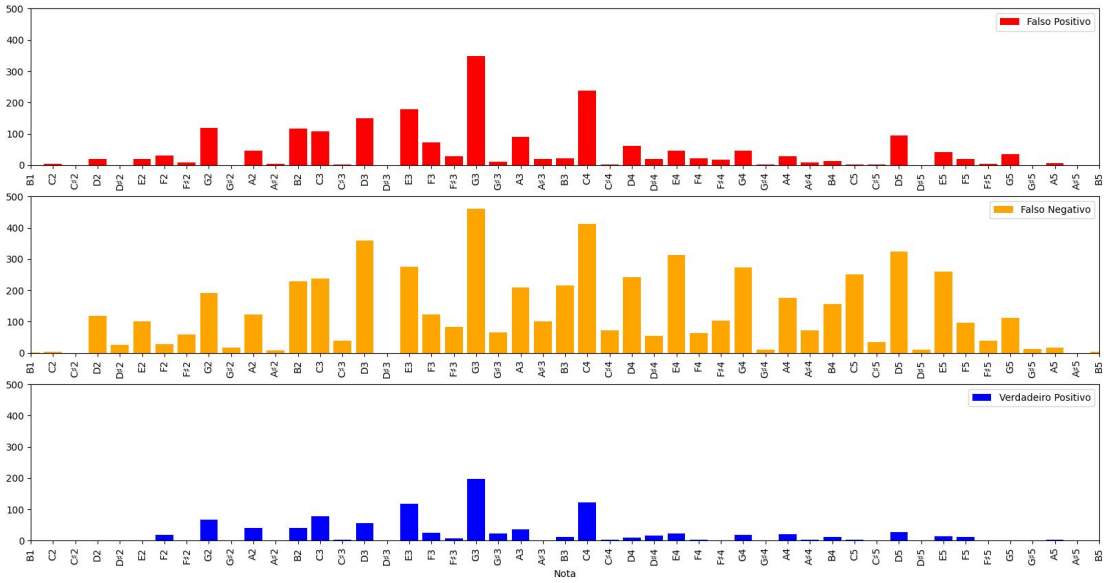
As figuras 6.13, 6.14 e 6.15 apresentam os histogramas deste experimento, que assim como o experimento anterior, demonstram a mesma tendência de diminuição dos falsos positivos e aumento dos falsos negativos conforme o treinamento avança.

Figura 6.13: Histograma - Experimento #04 - 3300



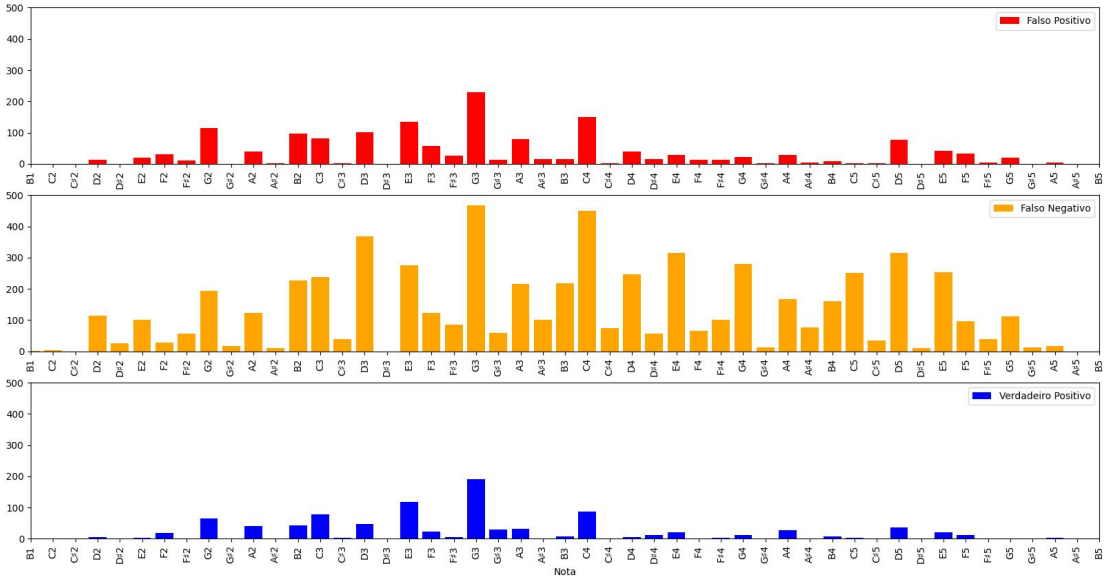
Fonte: Autoria Própria

Figura 6.14: Histograma - Experimento #04 - 6600



Fonte: Autoria Própria

Figura 6.15: Histograma - Experimento #04 - 9900



Fonte: Autoria Própria

	3300 Épocas			6600 Épocas			9900 Épocas		
Música	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
Primeira Valsa	0.336280	0.302550	**0.378490**	**0.380950**	0.417060	0.350600	0.322580	**0.427630**	0.258960
Baião	**0.169230**	0.236050	**0.131890**	0.166950	**0.288240**	0.117510	0.119850	0.273500	0.076740
feira de mangaio	**0.209320**	0.357140	**0.148040**	0.199130	0.416410	0.130850	0.198220	**0.439340**	0.127980
mineirinha	0.193010	0.214810	0.175230	0.210330	0.270140	0.172210	**0.246210**	**0.329950**	**0.196370**
missioneiro	**0.066990**	**0.116020**	**0.047090**	0.043710	0.099210	0.028030	0.026470	0.084340	0.015700
noite feliz	0.321610	0.367820	**0.285710**	**0.356940**	0.488370	0.281250	0.338370	**0.523360**	0.250000
o velhinho do realejo	0.312090	0.387980	**0.261030**	0.305760	0.480310	0.224260	**0.315510**	**0.578430**	0.216910
o relógio bateu 3 horas	0.234000	0.240910	**0.227470**	0.252010	0.335710	0.201720	**0.281690**	**0.409840**	0.214590
olha para o céu	**0.218560**	0.410420	**0.148940**	0.184240	0.468600	0.114660	0.175270	**0.497240**	0.106380
pagode russo	**0.103630**	**0.142250**	**0.081510**	0.073130	0.121470	0.052310	0.056390	0.123970	0.036500
parabéns	0.306570	0.362070	**0.265820**	**0.347110**	0.500000	**0.265820**	0.327590	**0.513510**	0.240510
sabia	**0.183670**	0.245050	**0.146880**	0.152690	0.277340	0.105340	0.158370	**0.333330**	0.103860
sanfoninha de ouro	0.381430	0.432330	0.341250	0.456140	0.557940	0.385760	**0.520150**	**0.679430**	**0.421360**
uma festa no ceu	0.356500	0.408300	0.316350	0.405360	0.540180	0.324400	**0.446370**	**0.629270**	**0.345840**
velhos tempos	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**

Tabela 6.6: Resultados do experimento #04

	3300 Épocas			6600 Épocas			9900 Épocas		
Música	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
baião	-0.10390	+0.03778	-0.30696	-0.10618	+0.08997	-0.32134	-0.15328	+0.07523	-0.36211
feira de mangaio	-0.00433	+0.14656	-0.06877	-0.01452	+0.20583	-0.08596	-0.01543	+0.22876	-0.08883
primeira valsa	+0.15912	+0.17734	+0.07570	+0.20379	+0.29185	+0.04781	+0.14542	+0.30242	-0.04383
mineirinha	+0.05213	+0.09412	+0.00605	+0.06945	+0.14945	+0.00303	+0.10533	+0.20926	+0.02719
missioneiro	-0.07404	-0.00601	-0.11995	-0.09732	-0.02282	-0.13901	-0.11456	-0.03769	-0.15134
noite feliz	-0.02783	+0.06846	-0.13393	+0.00750	+0.18901	-0.13839	-0.01107	+0.22400	-0.16964
o relógio bateu 3 horas	+0.05903	+0.11346	-0.05150	+0.07704	+0.20826	-0.07725	+0.10672	+0.28239	-0.06438
o velhinho do realejo	+0.11765	+0.23173	+0.00368	+0.11132	+0.32406	-0.03309	+0.12107	+0.42218	-0.04044
olha para o céu	-0.07577	+0.15119	-0.19149	-0.11009	+0.20937	-0.22577	-0.11906	+0.23801	-0.23405
pagode russo	-0.01819	+0.03046	-0.05231	-0.04869	+0.00968	-0.08151	-0.06543	+0.01218	-0.09732
parabéns	+0.04521	+0.12496	-0.02532	+0.08575	+0.26289	-0.02532	+0.06623	+0.27640	-0.05063
sabia	-0.13513	-0.00016	-0.30861	-0.16611	+0.03213	-0.35015	-0.16043	+0.08812	-0.35163
sanfoninha de ouro	+0.13816	+0.23156	+0.03264	+0.21287	+0.35717	+0.07715	+0.27688	+0.47866	+0.11275
uma festa no céu	+0.15439	+0.24192	+0.05898	+0.20325	+0.37380	+0.06703	+0.24426	+0.46289	+0.08847
velhos tempos	-0.00698	-0.00583	-0.00870	-0.00698	-0.00583	-0.00870	-0.00698	-0.00583	-0.00870

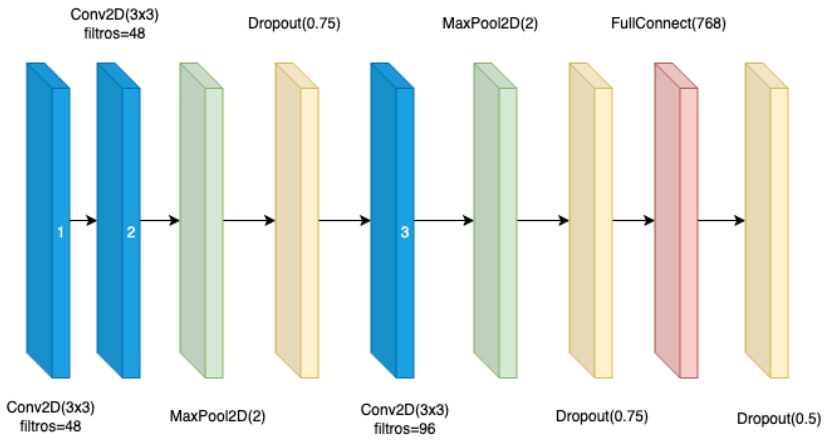
Tabela 6.7: Resultados do experimento #04 - Comparação com Baseline

6.5 Experimento #05: Todas as camadas congeladas

O objetivo deste experimento é avaliar o modelo aplicando a técnica de transferência de aprendizado chamada transferência de representação de características como explicado na seção 3.2.2. Neste experimento foram congeladas todas as camadas convolucionais como demonstrado pela figura 6.12 e realizado o treinamento por 3300, 6600 e 9900 épocas.

A tabela 6.8 demonstra os resultados deste experimento, assim como nos demais experimentos a música "sanfoninha de ouro" obteve o melhor desempenho, enquanto a música "velhos tempos" obteve o pior desempenho na transcrição.

Figura 6.16: Pilha Convolucional - Todas camadas convolucionais congeladas



Fonte: Autoria Própria

Com 3300 épocas houve uma melhora em f1 de 8 músicas e uma piora em 7 músicas, precision teve uma melhora em 13 músicas e uma piora em 2 e recall apresentou uma melhora em 3 músicas e uma piora em 12, apresentando uma perca média de ≈ 0,04% em f1, um ganho médio de ≈ 10,8% em precision e uma perca média de ≈ 10,1% em recall.

Com 6600 épocas houve uma melhora em f1 de 7 músicas e uma piora em 8 músicas, precision teve uma melhora em 13 músicas e uma piora em duas e recall apresentou uma melhora em 4 músicas e uma piora em 11, apresentando um ganho médio de ≈ 2,4% em f1, um ganho médio de ≈ 17% em precision e uma perca média de ≈ 9% em recall.

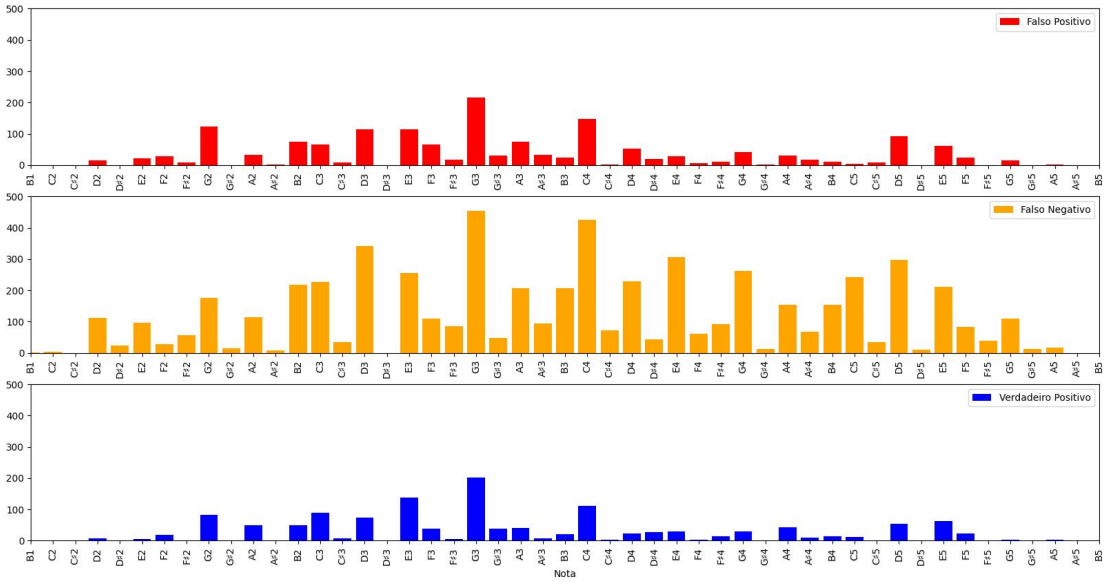
Com 9900 épocas houve uma melhora em f1 de 7 músicas e uma piora em 8 músicas, precision teve uma melhora em 13 músicas e uma piora em duas e

recall apresentou uma melhora em 3 músicas e uma piora em 12, apresentando um ganho médio de ≈ 2,4% em f1, um ganho médio de ≈ 21% em precision e uma perca média de ≈ 9,9% em recall.

A tabela 6.9 demonstra a diferença por música em comparação ao experimento baseline.

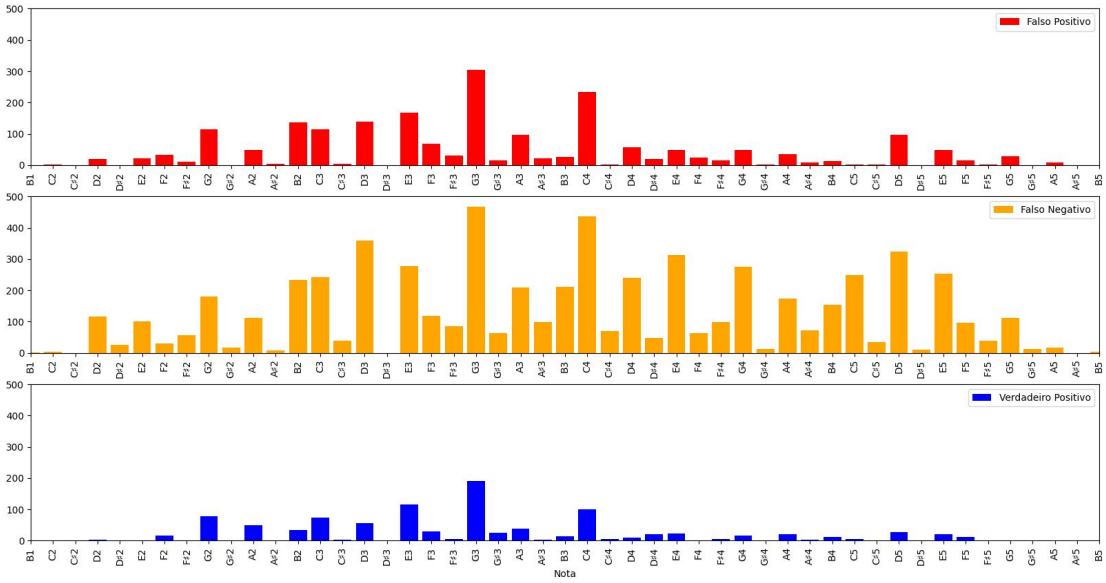
As figuras 6.17, 6.18 e 6.19 apresentam os histogramas deste experimento, que assim como os demais experimentos, demonstram a mesma tendência de diminuição dos falsos positivos e aumento dos falsos negativos conforme o treinamento avança.

Figura 6.17: Histograma - Experimento #05 - 3300



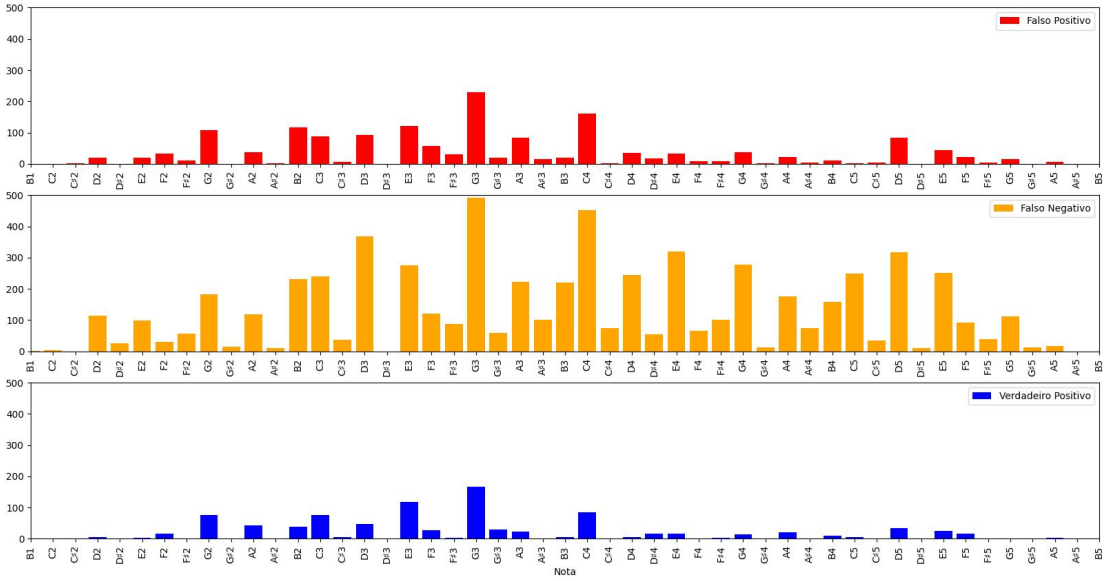
Fonte: Autoria Própria

Figura 6.18: Histograma - Experimento #05 - 6600



Fonte: Autoria Própria

Figura 6.19: Histograma - Experimento #05 - 9900



Fonte: Autoria Própria

	3300 Épocas			6600 Épocas			9900 Épocas		
Música	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
Primeira Valsa	0.306840	0.290000	**0.330000**	**0.359650**	0.400000	0.326690	0.314770	**0.401230**	0.258960
baiao	0.147200	0.220000	**0.110000**	**0.158730**	**0.300000**	0.107910	0.129390	0.282260	0.083930
feira de mangaio	0.172190	0.330000	0.120000	**0.204910**	0.418880	**0.135630**	0.195780	**0.462630**	0.124160
mineirinha	0.179360	0.200000	0.160000	**0.203570**	**0.248910**	**0.172210**	**0.27787**	**0.35185**	**0.22960**
missioneiro	**0.064620**	**0.120000**	**0.040000**	0.037670	0.094170	0.023540	0.026540	0.085890	0.015700
noite feliz	**0.351350**	0.450000	**0.290000**	0.334310	0.487180	0.254460	0.344830	**0.578950**	0.245540
o velhinho do realejo	0.312780	0.390000	**0.260000**	0.319200	0.496120	0.235290	**0.341460**	**0.649480**	0.231620
o relógio bateu 3 horas	0.221660	0.270000	0.190000	**0.278550**	**0.396830**	**0.214590**	0.262110	0.389830	0.197420
olha para o céu	0.147230	0.390000	0.090000	**0.187970**	0.458720	**0.118200**	0.157020	**0.462430**	0.094560
pagode russo	0.083410	0.150000	0.060000	**0.090280**	**0.157580**	**0.063260**	0.064520	0.146550	0.041360
parabéns	0.267720	0.350000	0.220000	0.305080	0.461540	0.227850	**0.319330**	**0.475000**	**0.240510**
sabiá	0.127610	0.250000	0.090000	**0.174350**	0.268520	**0.129080**	0.159480	**0.291340**	0.109790
sanfoninha de ouro	0.332760	0.400000	0.280000	0.432730	0.558690	0.353120	**0.530300**	**0.732980**	**0.415430**
uma festa no céu	0.329600	0.410000	0.280000	0.401290	0.506120	**0.332440**	**0.428570**	**0.611940**	0.329760
velhos tempos	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**

Tabela 6.8: Resultados do experimento #05 - Todas as camadadas convolucionais congeladas

	3300 Épocas			6600 Épocas			9900 Épocas		
Música	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
baião	-0.12593	+0.02173	-0.32885	-0.11440	+0.10173	-0.33094	-0.14374	+0.08399	-0.35492
feira de mangaio	-0.04146	+0.11942	-0.09681	-0.00874	+0.20830	-0.08118	-0.01787	+0.25205	-0.09265
primeira valsa	+0.12968	+0.16479	+0.02721	+0.18249	+0.27479	+0.02390	+0.13761	+0.27602	-0.04383
mineirinha	+0.03848	+0.07931	-0.00918	+0.06269	+0.12822	+0.00303	+0.13699	+0.23117	+0.06042
missioneiro	-0.07641	-0.00203	-0.12704	-0.10336	-0.02786	-0.14350	-0.11449	-0.03614	-0.15134
noite feliz	+0.00191	+0.15064	-0.12964	-0.01513	+0.18782	-0.16518	-0.00461	+0.27959	-0.17410
o relógio bateu 3 horas	+0.04669	+0.14255	-0.08897	+0.10358	+0.26938	-0.06438	+0.08714	+0.26238	-0.08155
o velhinho do realejo	+0.11834	+0.23375	+0.00265	+0.12476	+0.33987	-0.02206	+0.14702	+0.49323	-0.02573
olha para o céu	-0.14710	+0.13077	-0.25043	-0.10636	+0.19949	-0.22223	-0.13731	+0.20320	-0.24587
pagode russo	-0.03841	+0.03821	-0.07382	-0.03154	+0.04579	-0.07056	-0.05730	+0.03476	-0.09246
parabéns	+0.00636	+0.11289	-0.07114	+0.04372	+0.22443	-0.06329	+0.05797	+0.23789	-0.05063
sabiá	-0.19119	+0.00479	-0.36549	-0.14445	+0.02331	-0.32641	-0.15932	+0.04613	-0.34570
sanfoninha de ouro	+0.08949	+0.19923	-0.02861	+0.18946	+0.35792	+0.04451	+0.28703	+0.53221	+0.10682
uma festa no céu	+0.12749	+0.24362	+0.02263	+0.19918	+0.33974	+0.07507	+0.22646	+0.44556	+0.07239
velhos tempos	-0.00698	-0.00583	-0.00870	-0.00698	-0.00583	-0.00870	-0.00698	-0.00583	-0.00870

Tabela 6.9: Resultados do experimento #05 - Comparação com Baseline

6.6 Experimento #06: Piano

O objetivo deste experimento é demonstrar a perca de conhecimento do problema original e o impacto de se congelar diferentes camadas convolucionais durante o treinamento dos experimentos #02, #03, #04 e #05.

Para as métricas do piano foram utilizadas todas 125 músicas classificadas como teste durante o treinamento do modelo original. A tabela 6.10 apresenta a média das métricas destas músicas antes de ser realizado qualquer ajuste no modelo, enquanto a tabela 6.11 apresenta a média das métricas dos demais experimentos.

Ao analisar a tabela 6.11 é possível notar uma perca consistente do recall conforme o modelo é treinado por mais épocas e que o número de camadas convolucionais congeladas é proporcional a perca de conhecimento da base original.

f1	precision	recall
0.947435	0.979251	0.918295

Tabela 6.10: Resultado da base de teste piano - experimento #01 Baseline

	3300 Épocas			6600 Épocas			9900 Épocas		
Experimento	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
#02 - Fine Tunning	0.774510	0.945730	0.664813	0.670266	0.950072	0.527686	0.561723	0.949795	0.408274
#03 - 1 Camada	0.777847	0.949611	0.667377	0.671972	0.949155	0.529649	0.556988	0.946175	0.402460
#04 - 2 Camadas	0.767120	0.948356	0.651909	0.663774	0.949030	0.519659	0.540767	0.946201	0.386141
#05 - 3 Camadas	0.774732	0.949924	0.662743	0.654136	0.949103	0.508163	0.542023	0.944102	0.389009

Tabela 6.11: Resultado da base de teste piano

6.7 Experimento #07: Variação de registros utilizando apenas teclado

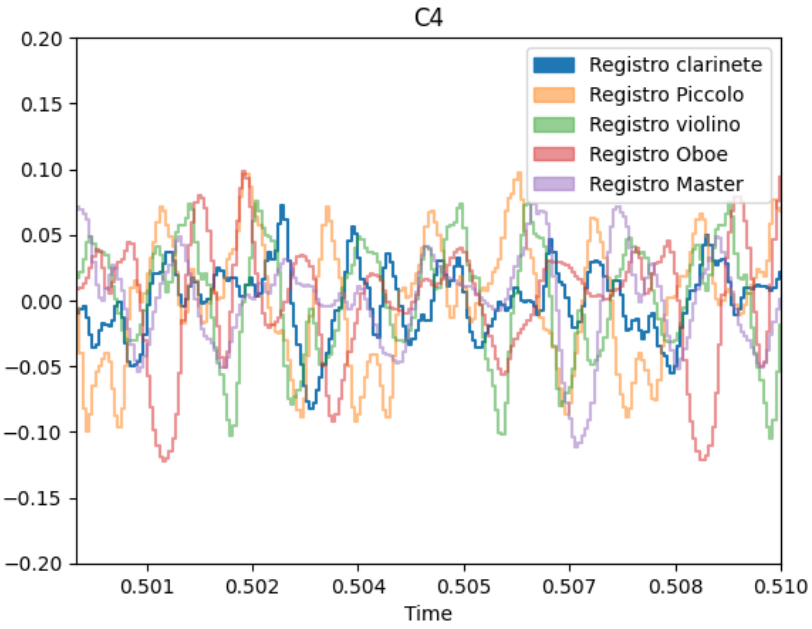
Uma característica presente na maioria das Sanfonas são seus diferentes registros, como mencionado na seção 2.3 uma mudança de registro significa uma mudança no timbre do som.

O objetivo deste experimento é demonstrar o impacto que estes diferentes registros tem nas métricas do modelo.

A tabela 6.12 demonstra as métricas dos áudios gravados utilizando apenas o teclado da Sanfona para cada um dos registros disponíveis no dataset utilizando a mesma configuração do experimento #01 - Baseline.

Analisando as métricas observando apenas as métricas do experimento #01 - Baseline, tabela 6.12, é possível notar que em ambas as músicas os baixos Violino e Master foram os registros com as melhores transcrições, os timbres da Sanfona apesar de terem a mesma frequência base produzem ondas sonoras de formatos diferentes como demonstrado pela imagem 6.20 impactando diretamente na qualidade da transcrição.

Figura 6.20: Ondas da nota C4 (teclado) - Diferentes Registros



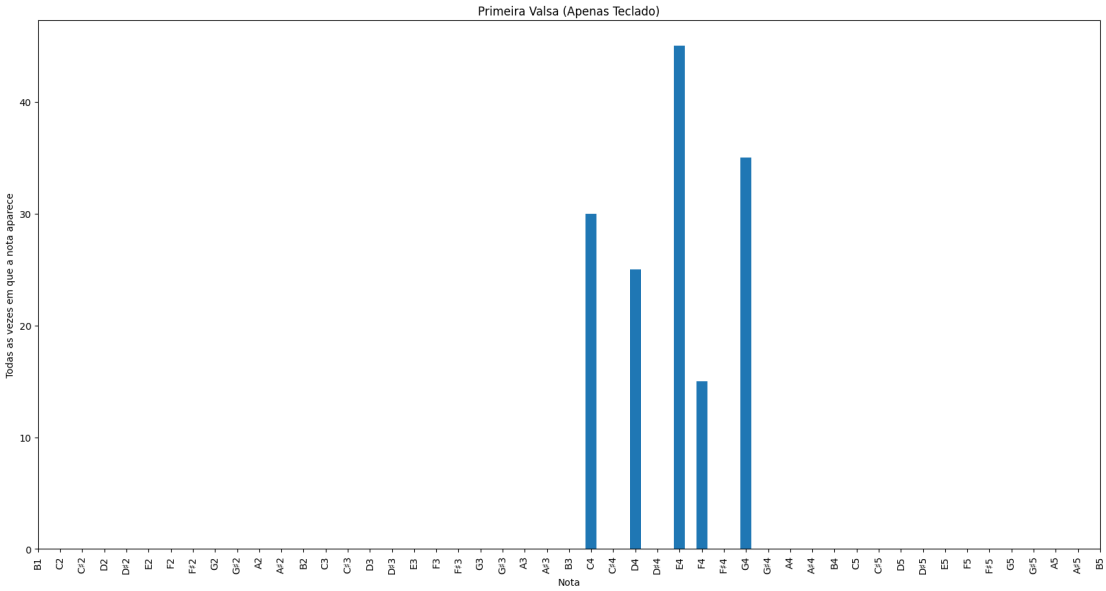
Fonte: Autoria Própria

As tabelas 6.13 e 6.14 demonstram as métricas das músicas Primeira Valsa e

O Relógio bateu três horas utilizando apenas o teclado e com diferentes timbres.

As figuras 6.21 e 6.22 demonstram as notas presentes na avaliação deste experimento, onde é possível notas haver apenas cinco notas diferentes sendo analisadas em diferentes registros.

Figura 6.21: Primeira Valsa (configuração apenas teclado)



Fonte: Autoria Própria

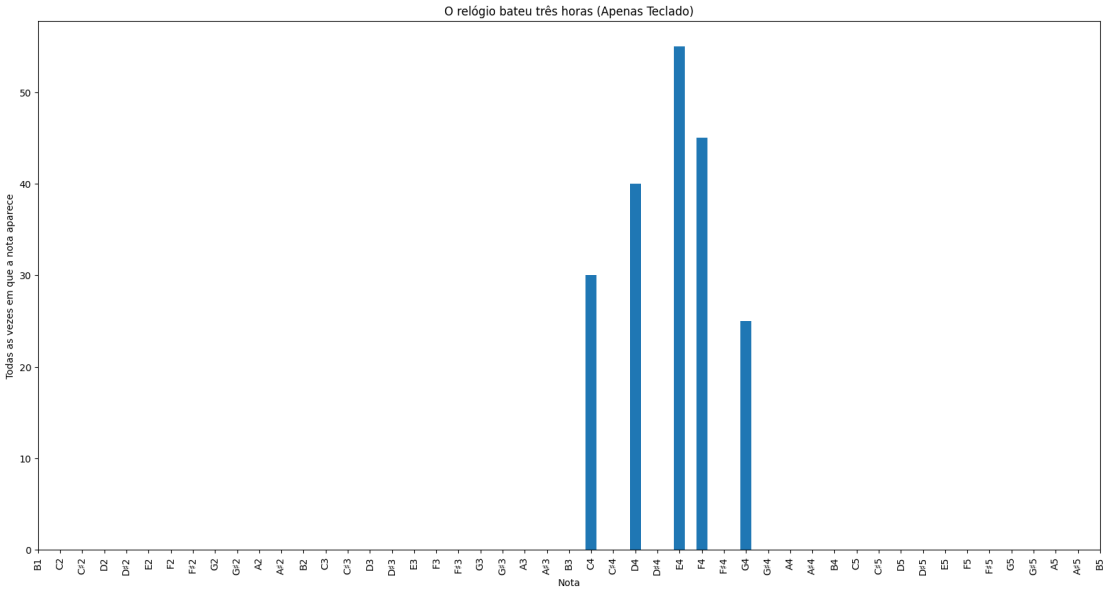
Os resultados demonstram que apenas utilizando o teclado da Sanfona e variando os registros o modelo não conseguiu prever as notas corretamente, resultado que já esperado, por existirem apenas duas músicas na base de dados que apresentam as variações de registros no teclado e sem a inclusão dos baixos (Primeira Valsa e O relógio bateu três horas).

As figuras 6.23, 6.24 e 6.25 mostram a evolução das métricas do modelo para a música Primeira Valsa utilizando apenas o teclado (incluindo a variação dos registros) durante a execução do experimento #02.

Vale ressaltar que conforme se aumenta a quantidade de épocas de treinamento o modelo tende a diminuir os falsos positivos que se concentram principalmente nas notas E5, D5, B5, porém, tende a diminuir os verdadeiros positivos. Neste experimento o modelo conseguiu identificar parcialmente apenas as notas E4 e G4.

Já as figuras 6.26, 6.27, 6.28, 6.29, 6.30, 6.31, 6.32, 6.33 e 6.34 demonstram que a quantidade de camadas convolucionais congeladas é inversamente proporcional

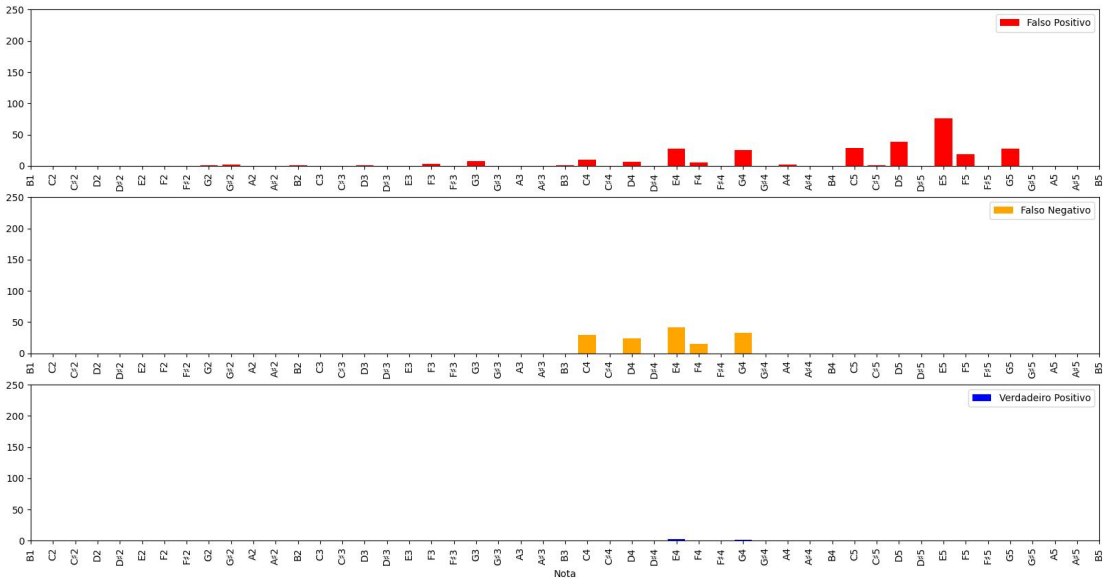
Figura 6.22: O relógio bateu três horas (configuração apenas teclado)



Fonte: Autoria Própria

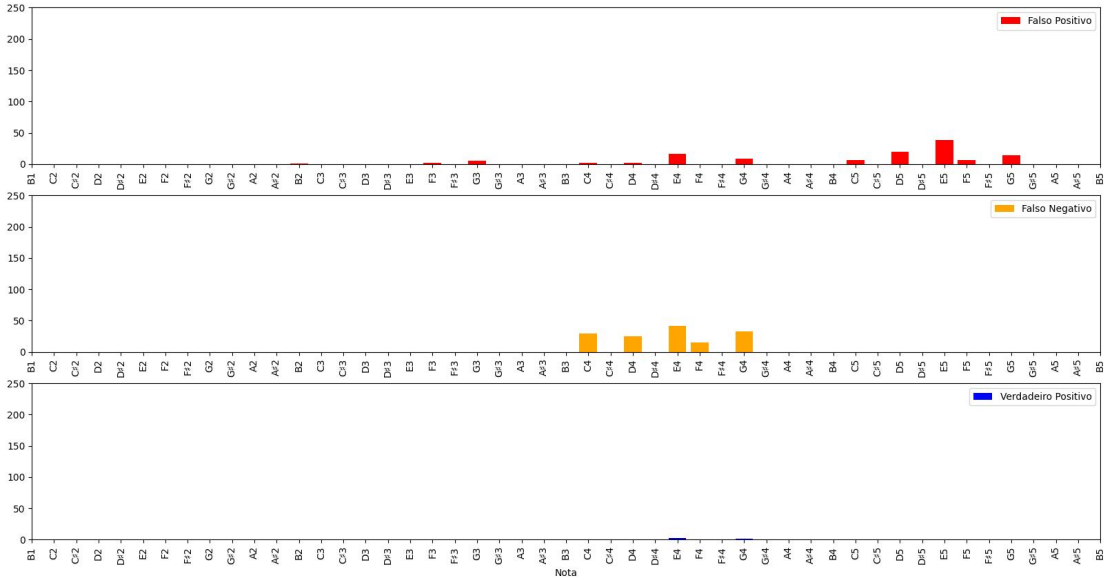
a quantidade de épocas necessárias no treinamento para atingir os mesmos resultados para as músicas que contém apenas o teclado da Sanfona.

Figura 6.23: Histograma Primeira Valsa (Apenas Teclado) - Experimento #02 - 3300



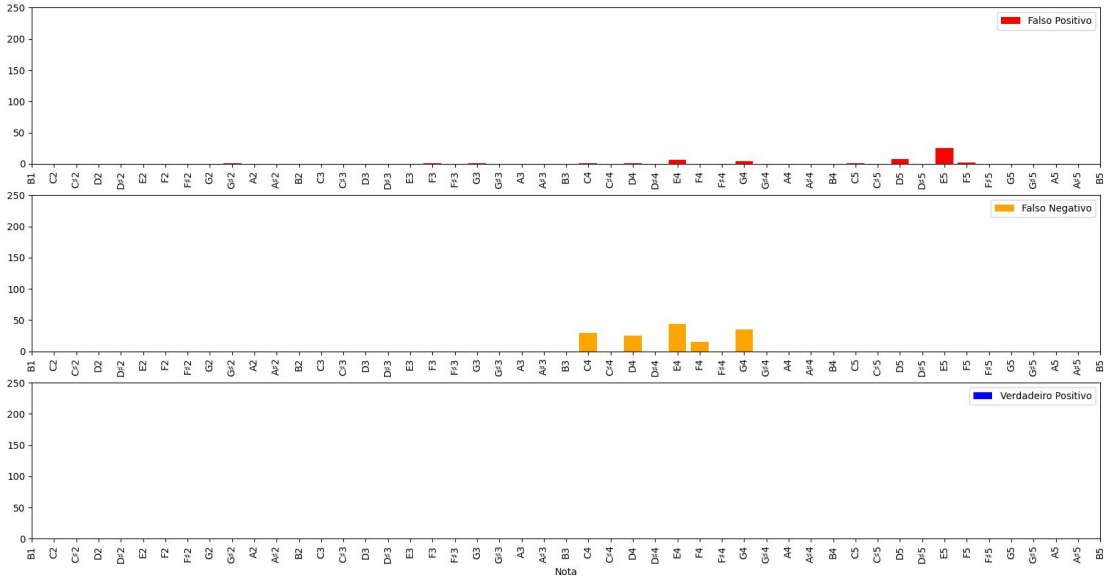
Fonte: Autoria Própria

Figura 6.24: Histograma Primeira Valsa (Apenas Teclado) - Experimento #02 - 6600



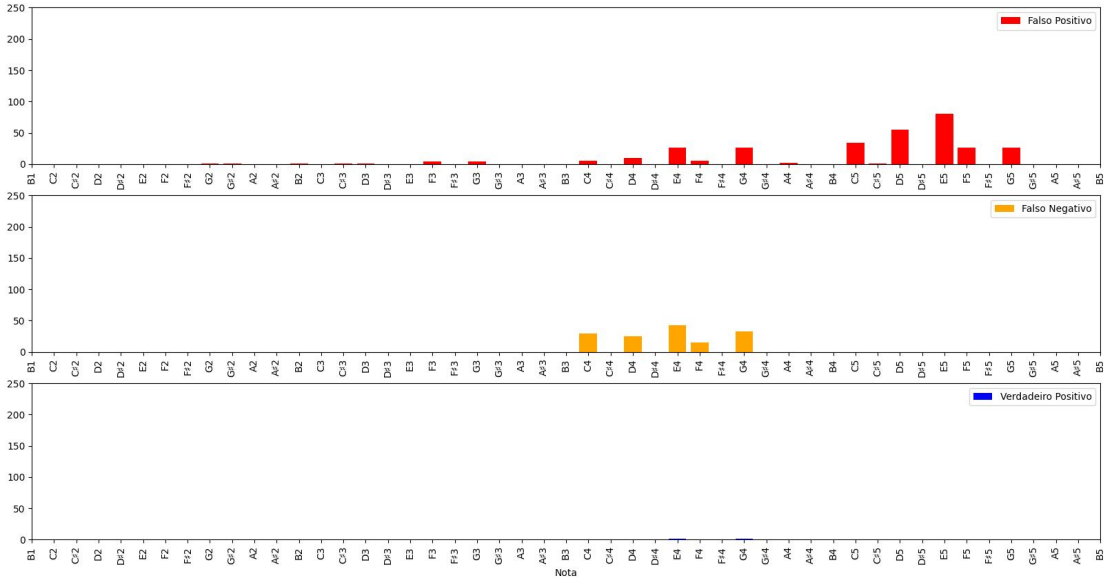
Fonte: Autoria Própria

Figura 6.25: Histograma Primeira Valsa (Apenas Teclado) - Experimento #02 - 9900



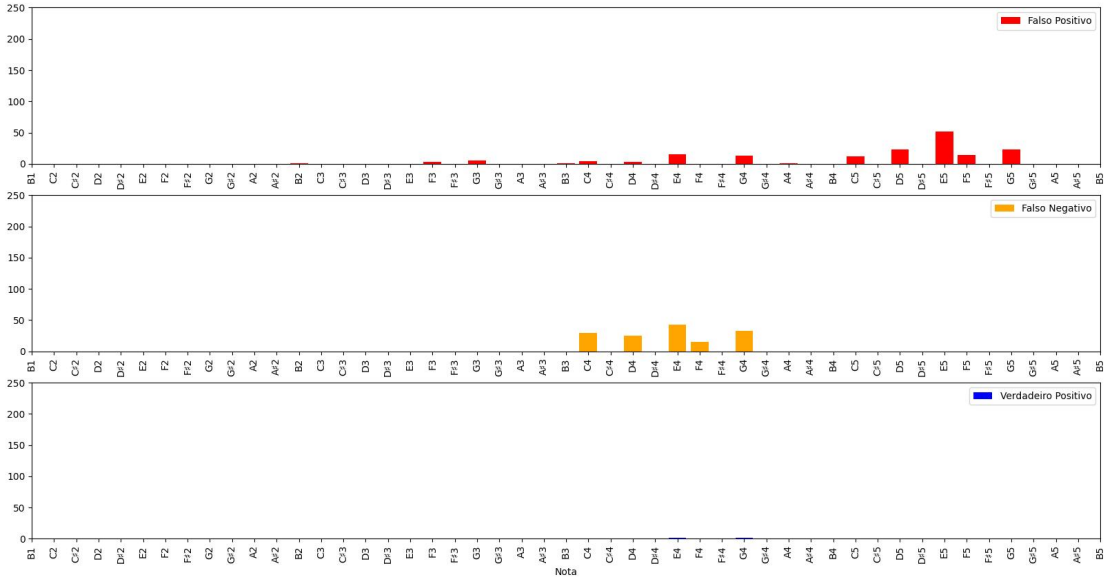
Fonte: Autoria Própria

Figura 6.26: Histograma Primeira Valsa (Apenas Teclado) - Experimento #03 - 3300



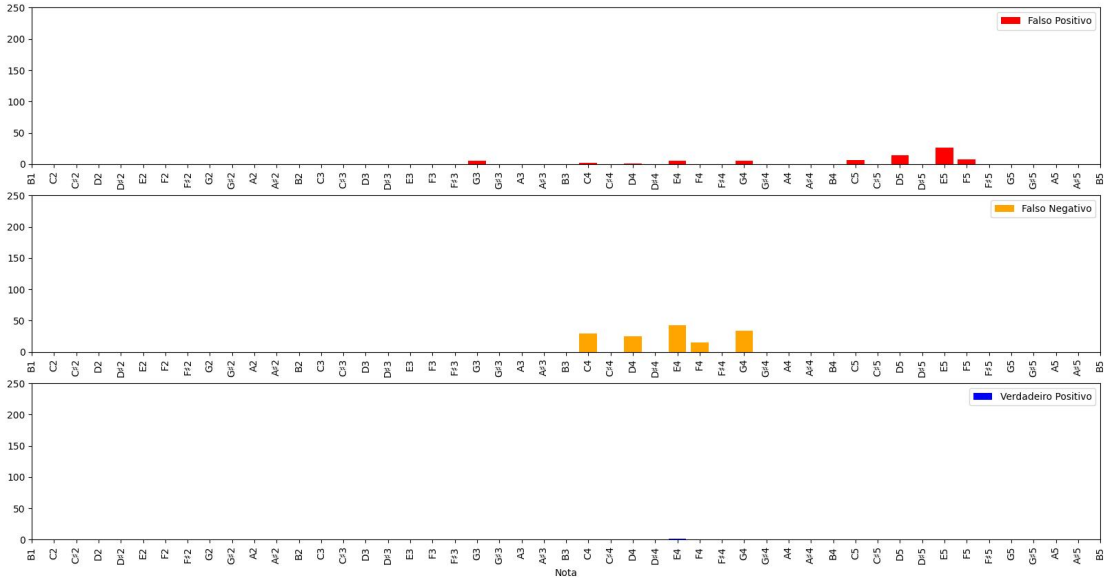
Fonte: Autoria Própria

Figura 6.27: Histograma Primeira Valsa (Apenas Teclado) - Experimento #03 - 6600



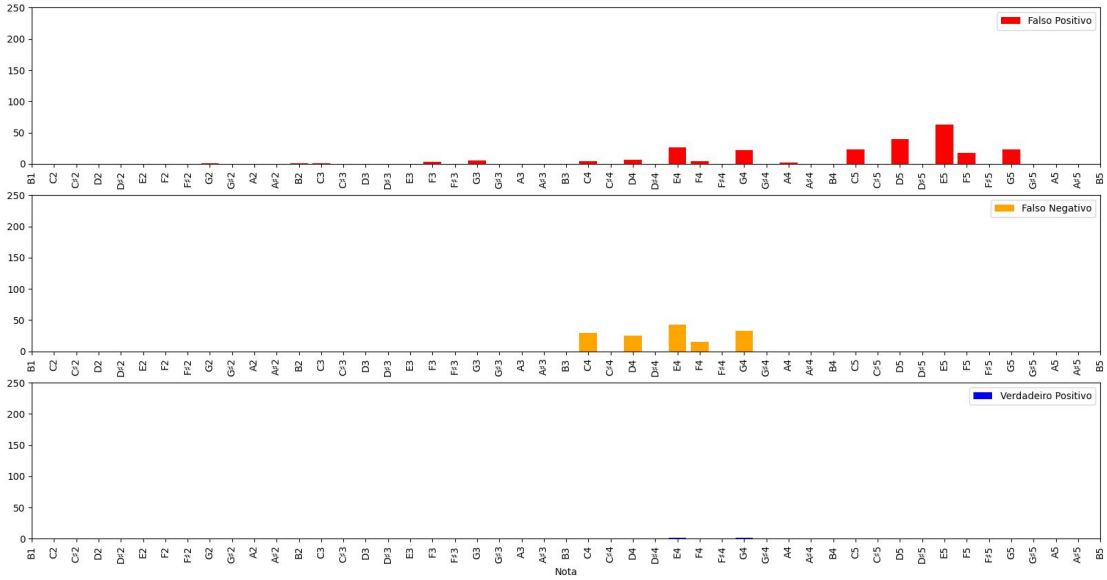
Fonte: Autoria Própria

Figura 6.28: Histograma Primeira Valsa (Apenas Teclado) - Experimento #03 - 9900



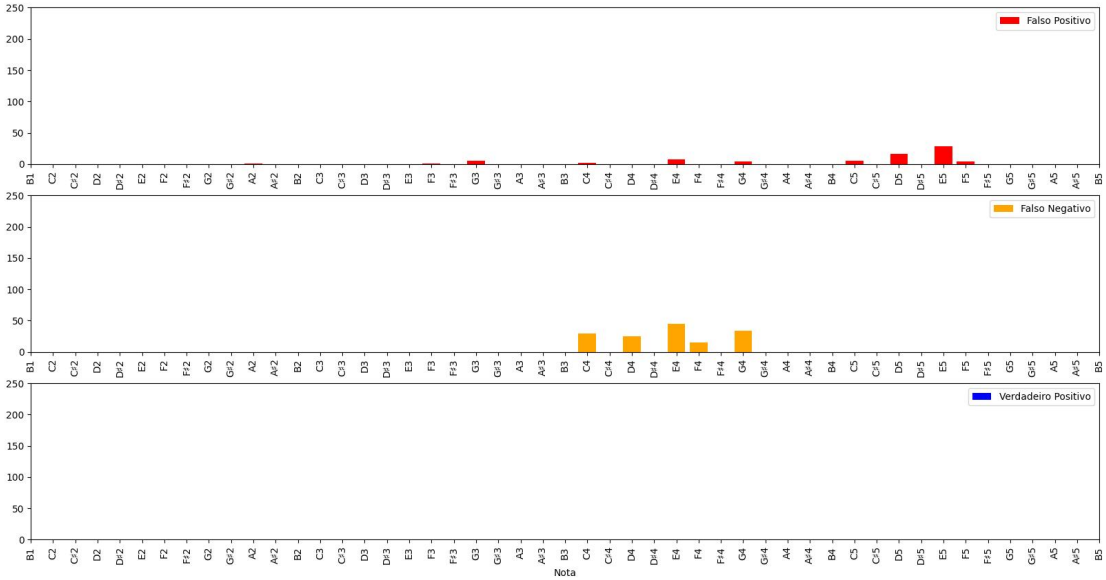
Fonte: Autoria Própria

Figura 6.29: Histograma Primeira Valsa (Apenas Teclado) - Experimento #04 - 3300



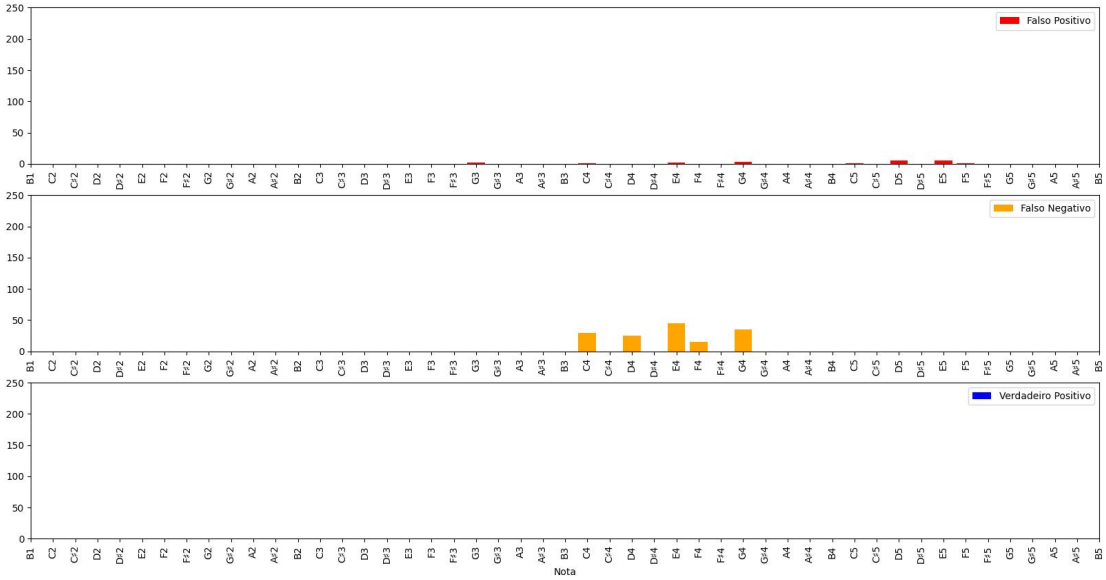
Fonte: Autoria Própria

Figura 6.30: Histograma Primeira Valsa (Apenas Teclado) - Experimento #04 - 6600



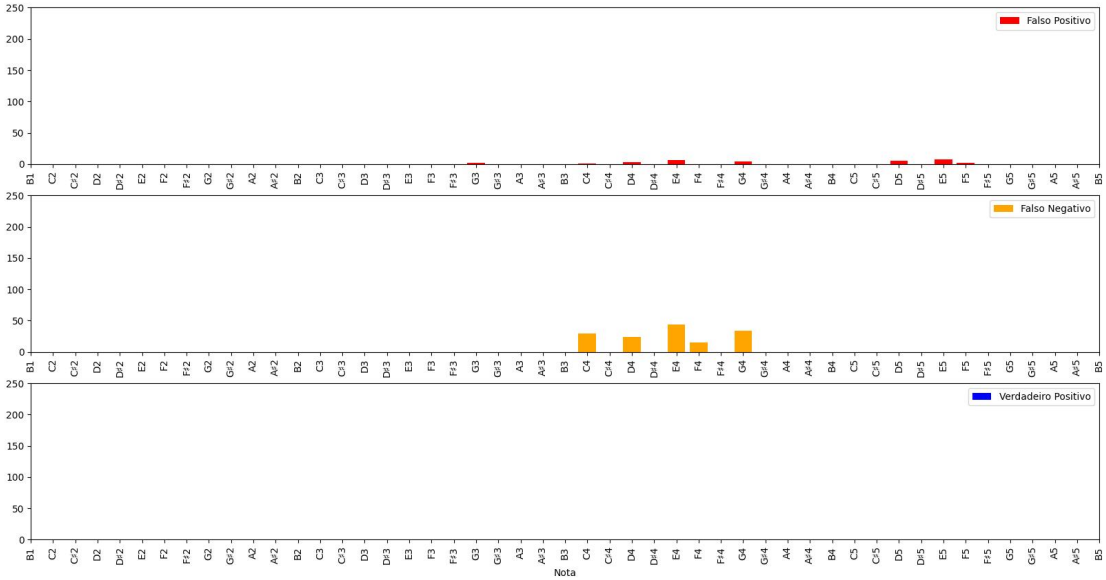
Fonte: Autoria Própria

Figura 6.31: Histograma Primeira Valsa (Apenas Teclado) - Experimento #04 - 9900



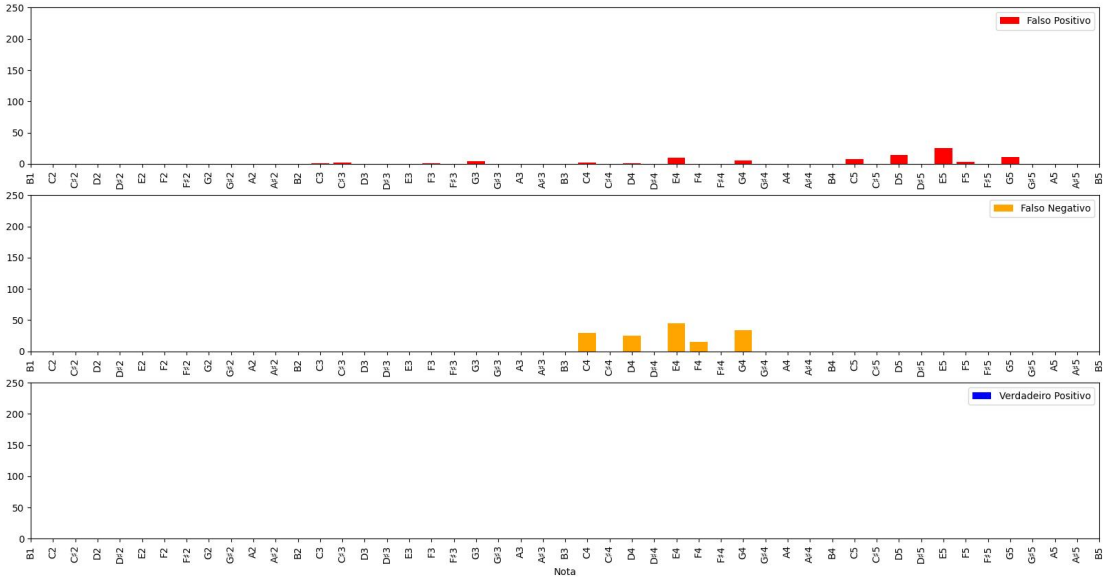
Fonte: Autoria Própria

Figura 6.32: Histograma Primeira Valsa (Apenas Teclado) - Experimento #05 - 3300



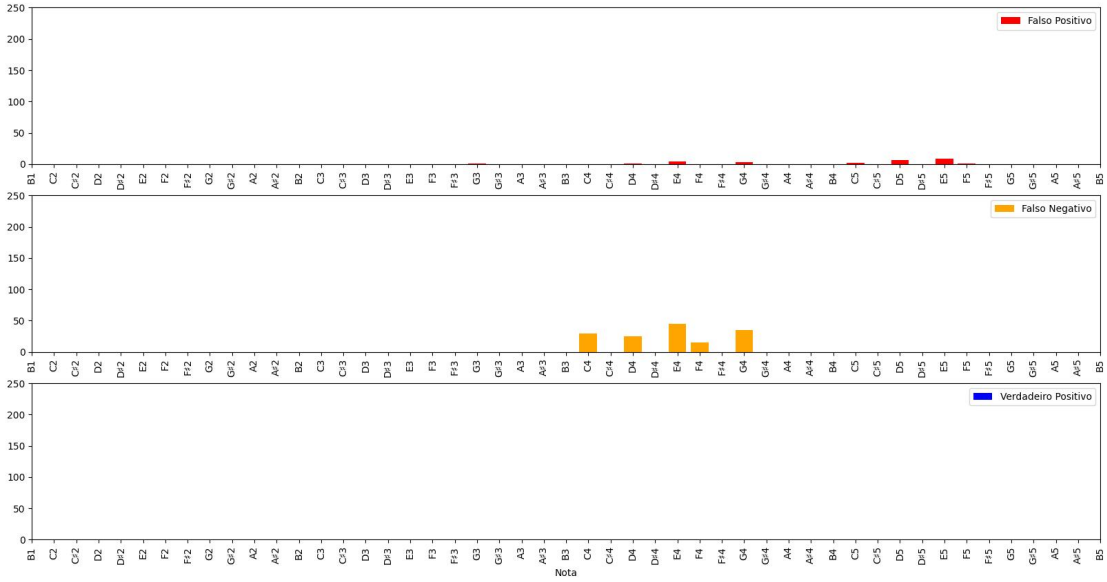
Fonte: Autoria Própria

Figura 6.33: Histograma Primeira Valsa (Apenas Teclado) - Experimento #05 - 6600



Fonte: Autoria Própria

Figura 6.34: Histograma Primeira Valsa (Apenas Teclado) - Experimento #05 - 9900



Fonte: Autoria Própria

Música	**Timbre**	**f1-score**	**Precision**	**Recall**
Primeira valsa	Clarinete	0,0	0,0	0,0
Primeira valsa	Piccolo	0,125	0,1	0,16666
Primeira valsa	Violino	**0,428570**	0,30882	0,7
Primeira valsa	Oboe	0,0	0,0	0,0
Primeira valsa	Master	0,04347	0,03225	0,06666
O relógio bateu 3 horas	Clarinete	0,0	0,0	0,0
O relógio bateu 3 horas	Piccolo	0,0	0,0	0,0
O relógio bateu 3 horas	Violino	0,01694	0,012658	0,02564
O relógio bateu 3 horas	Oboe	0,0	0,0	0,0
O relógio bateu 3 horas	Master	**0,12121**	0,07936	0,25641

Tabela 6.12: Variação de timbres - Baseline. Configuração: Apenas Teclado

		3300 Épocas			6600 Épocas			9900 Épocas		
Experimento	**Timbre**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
#02 - Fine Tunning	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#02 - Fine Tunning	Piccolo	0.057140	0.050000	**0.066670**	**0.076920**	**0.090910**	**0.066670**	0.000000	0.000000	0.000000
#02 - Fine Tunning	Violino	**0.081080**	0.068180	**0.100000**	0.076920	**0.090910**	0.066670	0.045450	0.071430	0.033330
#02 - Fine Tunning	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#02 - Fine Tunning	Master	0.033330	0.033330	**0.033330**	**0.040820**	**0.052630**	**0.033330**	0.000000	0.000000	0.000000
#03 - 1 Camada Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Piccolo	0.061540	0.057140	**0.066670**	**0.072730**	**0.080000**	**0.066670**	0.046510	0.076920	0.033330
#03 - 1 Camada Cong.	Violino	0.033900	0.034480	**0.033330**	0.038460	0.045450	**0.033330**	**0.055560**	**0.166670**	**0.033330**
#03 - 1 Camada Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Master	0.031750	0.030300	**0.033330**	0.041670	0.055560	**0.033330**	**0.045450**	**0.071430**	**0.033330**
#04 - 2 Camadas Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Piccolo	**0.064520**	**0.062500**	**0.066670**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#04 - 2 Camadas Cong.	Violino	**0.036360**	**0.040000**	**0.033330**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#04 - 2 Camadas Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Master	0.033900	0.034480	**0.033330**	**0.048780**	**0.090910**	**0.033330**	0.000000	0.000000	0.000000
#05 - T. Camadas Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#05 - T. Camadas Cong.	Piccolo	**0.090910**	**0.080000**	**0.100000**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#05 - T. Camadas Cong.	Violino	**0.113210**	**0.130000**	**0.100000**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#05 - T. Camadas Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#05 - T. Camadas Cong.	Master	0.036360	0.040000	0.030000	**0.046510**	**0.076920**	**0.033330**	0.000000	0.000000	0.000000

Tabela 6.13: Resultados Primeira Valsa - Variação dos registros. Configuração: Apenas teclado com diferentes Timbres

		3300 Épocas			6600 Épocas			9900 Épocas		
Experimento	**Timbre**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
#02 - Fine Tunning	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#02 - Fine Tunning	Piccolo	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#02 - Fine Tunning	Violino	**0.027400**	**0.029410**	**0.025640**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#02 - Fine Tunning	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#02 - Fine Tunning	Master	0.028170	0.031250	**0.025640**	0.000000	0.000000	0.000000	**0.043480**	**0.142860**	**0.025640**
#03 - 1 Camada Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Piccolo	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Violino	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Master	0.050000	0.048780	0.051280	**0.140350**	**0.222220**	**0.102560**	0.113210	0.214290	0.076920
#04 - 2 Camadas Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Piccolo	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Violino	**0.025970**	**0.026320**	**0.025640**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#04 - 2 Camadas Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Master	0.000000	0.000000	0.000000	0.039220	0.083330	**0.025640**	**0.044440**	**0.166670**	**0.025640**
#05 - T. Camadas Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#05 - T. Camadas Cong.	Piccolo	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#05 - T. Camadas Cong.	Violino	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#05 - T. Camadas Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#05 - T. Camadas Cong.	Master	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**

Tabela 6.14: Resultados O relógio bateu três horas - Variação dos registros. Configuração: Apenas teclado com diferentes Timbres

6.8 Experimento #08: Variação de registros utilizando teclado e baixos

Neste experimento foram avaliadas as mesmas interpretações do experimento da seção 6.7, porém incluindo o ritmo na mão esquerda (baixos), a tabela 6.15 mostra que para a música Primeira Valsa o registro com o melhor desempenho foi o master e o pior foi o clarinete, em contrapartida, da música o relógio bateu três horas onde o clarinete teve o melhor desempenho.

Para o problema apresentado, a inclusão dos baixos significa a inclusão de acordes na mão esquerda, apesar de ser apenas um único botão sendo pressionado pelo músico, podem ser emitidas três ou mais notas pelo instrumento simultaneamente.

Música	**Timbre**	**f1-score**	**Precision**	**Recall**
Primeira valsa	Clarinete	0,00494	0,00358	0,00796
Primeira valsa	Piccolo	0,08602	0,06490	0,12749
Primeira valsa	Violino	0,124220	0,09025	0,19920
Primeira valsa	Oboe	0,15365	0,11151	0,24701
Primeira valsa	Master	**0,17715**	0,12520	0,30278
O relógio bateu 3 horas	Clarinete	**0,20903**	0,15578	0,31759
O relógio bateu 3 horas	Piccolo	0,01164	0,00881	0,01716
O relógio bateu 3 horas	Violino	0,20081	0,14682	0,31759
O relógio bateu 3 horas	Oboe	0,0082079	0,00602	0,01287
O relógio bateu 3 horas	Master	0,17496	0,12745	0,27896

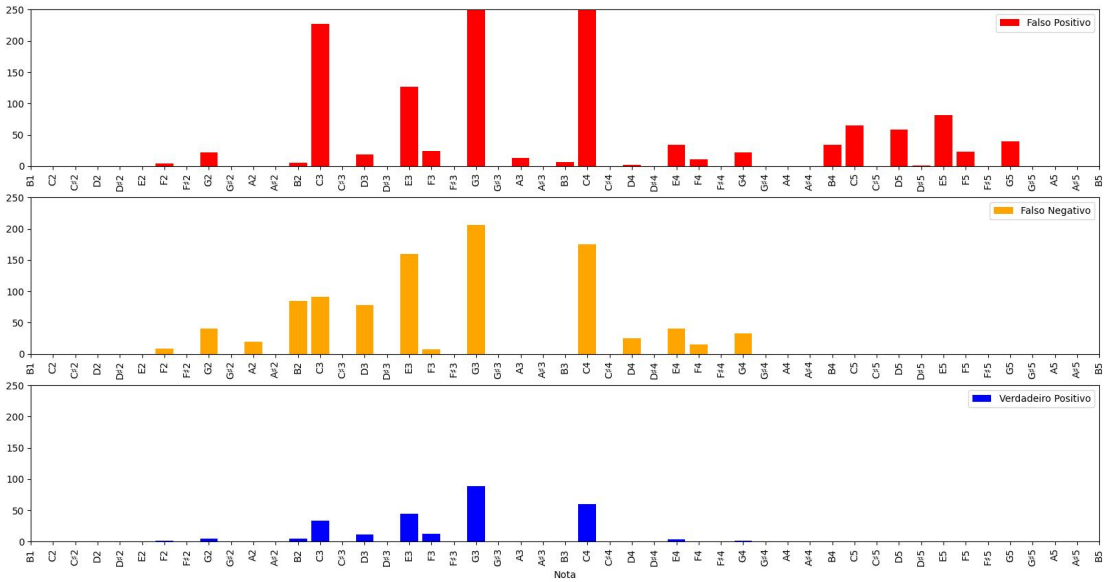
Tabela 6.15: Variação de timbres - Baseline. Configuração: Teclado e baixos

As tabelas 6.16 e 6.17 demonstram respectivamente as métricas das músicas "Primeira Valsa" e "O relógio bateu três horas" com interpretações utilizando tanto o teclado quanto os baixos do instrumento e incluindo a variação de registros do teclado.

As figuras 6.35, 6.36 e 6.37 apresentam o histograma de notas da música primeira valsa com a inclusão dos baixos. Nestas figuras é possível notar que as notas que o modelo tem maior facilidade em encontrar são E3, G3, C4, enquanto os erros estão concentrados nas notas C3, G3, C4.

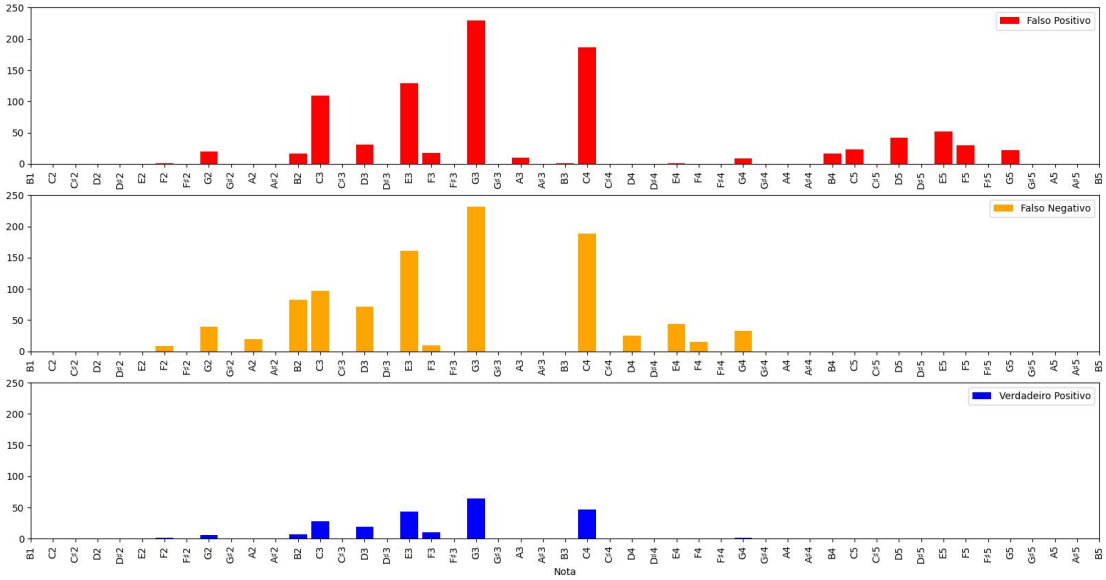
As figuras 6.38, 6.39, 6.40, 6.41, 6.42, 6.43, 6.44, 6.45 e 6.46 demonstram a mesma tendência de acerto.

Figura 6.35: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #02 - 3300



Fonte: Autoria Própria

Figura 6.36: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #02 - 6600



Fonte: Autoria Própria

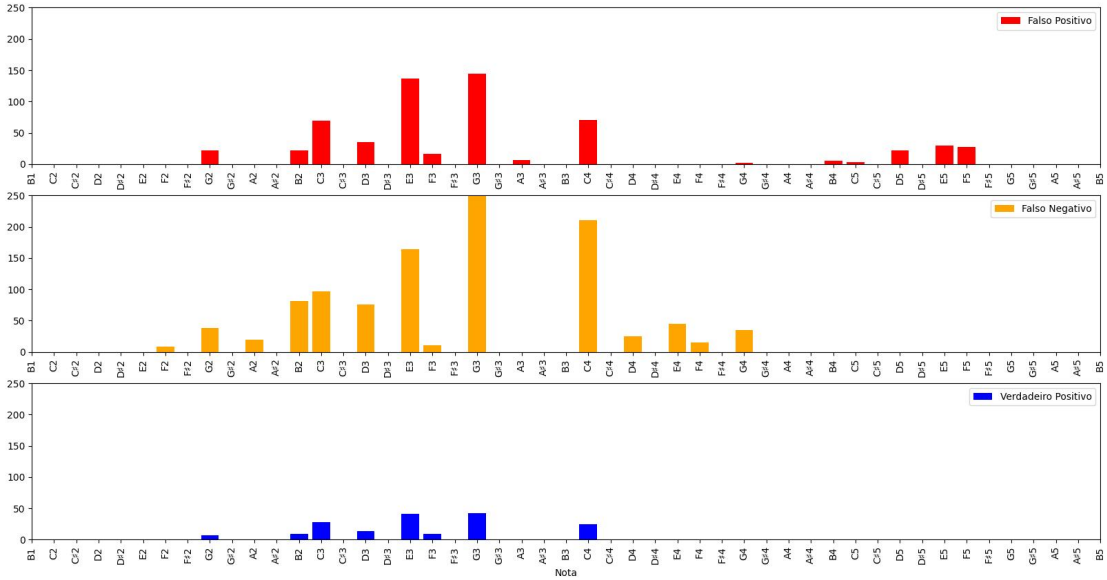
		3300 Épocas			6600 Épocas			9900 Épocas		
Experimento	**Timbre**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
#02 - Fine Tunning	Clarinete	**0.006130**	**0.004980**	**0.007970**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#02 - Fine Tunning	Piccolo	**0.104350**	0.092590	**0.119520**	0.099350	**0.108490**	0.091630	0.056700	0.080290	0.043820
#02 - Fine Tunning	Violino	0.188680	0.165660	**0.219120**	**0.214430**	0.222220	0.207170	0.209030	**0.258820**	0.175300
#02 - Fine Tunning	Oboe	0.277780	0.226700	**0.358570**	**0.284550**	0.290460	0.278880	0.271430	**0.337280**	0.227090
#02 - Fine Tunning	Master	0.321370	0.281440	**0.374500**	**0.349080**	0.360170	0.338650	0.313730	**0.407640**	0.254980
#03 - 1 Camada Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Piccolo	0.061540	0.057140	**0.066670**	**0.072730**	**0.080000**	**0.066670**	0.046510	0.076920	0.033330
#03 - 1 Camada Cong.	Violino	0.033900	0.034480	**0.033330**	0.038460	0.045450	**0.033330**	**0.055560**	**0.166670**	**0.033330**
#03 - 1 Camada Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Master	0.031750	0.030300	**0.033330**	0.041670	0.055560	**0.033330**	**0.045450**	**0.071430**	**0.033330**
#04 - 2 Camadas Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Piccolo	**0.064520**	**0.062500**	**0.066670**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#04 - 2 Camadas Cong.	Violino	**0.036360**	**0.040000**	**0.033330**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#04 - 2 Camadas Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Master	0.033900	0.034480	**0.033330**	**0.048780**	**0.090910**	**0.033330**	0.000000	0.000000	0.000000
#05 - T. Camadas Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#05 - T. Camadas Cong.	Piccolo	0.076230	0.070000	0.080000	**0.092920**	0.104480	**0.083670**	0.088890	**0.116880**	0.071710
#05 - T. Camadas Cong.	Violino	0.153260	0.150000	0.160000	**0.208070**	0.222730	**0.195220**	0.196640	**0.246990**	0.163350
#05 - T. Camadas Cong.	Oboe	0.222610	0.200000	0.250000	0.335430	0.353980	**0.318730**	**0.341920**	**0.414770**	0.290840
#05 - T. Camadas Cong.	Master	0.306840	0.290000	**0.330000**	**0.359650**	0.400000	0.326690	0.314770	**0.401230**	0.258960

Tabela 6.16: Resultados Primeira Valsa - Variação dos registros. Configuração: Teclado com diferentes Timbres + Baixos

		3300 Épocas			6600 Épocas			9900 Épocas		
Experimento	**Timbre**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**
#02 - Fine Tunning	Clarinete	0.327940	0.355000	0.304720	0.335920	0.422080	0.278970	**0.374030**	**0.473680**	**0.309010**
#02 - Fine Tunning	Piccolo	0.000000	0.000000	0.000000	**0.005290**	**0.006900**	**0.004290**	0.000000	0.000000	0.000000
#02 - Fine Tunning	Violino	0.262530	0.295700	0.236050	0.265580	0.360290	0.210300	**0.341460**	**0.463240**	**0.270390**
#02 - Fine Tunning	Oboe	0.012050	0.011320	0.012880	0.018780	0.020730	**0.017170**	**0.022600**	**0.033060**	**0.017170**
#02 - Fine Tunning	Master	0.247090	0.270410	0.227470	0.256410	0.381360	0.193130	**0.307260**	**0.440000**	**0.236050**
#03 - 1 Camada Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Piccolo	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Violino	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#03 - 1 Camada Cong.	Master	0.050000	0.048780	0.051280	**0.140350**	**0.222220**	**0.102560**	0.113210	0.214290	0.076920
#04 - 2 Camadas Cong.	Clarinete	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Piccolo	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Violino	**0.025970**	**0.026320**	**0.025640**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
#04 - 2 Camadas Cong.	Oboe	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
#04 - 2 Camadas Cong.	Master	0.000000	0.000000	0.000000	0.039220	0.083330	**0.025640**	**0.044440**	**0.166670**	**0.025640**
#05 - T. Camadas Cong.	Clarinete	0.263680	0.310000	0.230000	**0.306410**	0.436510	**0.236050**	0.276920	**0.489130**	0.193130
#05 - T. Camadas Cong.	Piccolo	0.004760	**0.010000**	0.000000	0.005330	0.007040	**0.004290**	**0.005760**	0.008770	**0.004290**
#05 - T. Camadas Cong.	Violino	0.199520	0.220000	0.180000	0.258950	0.361540	0.201720	**0.299710**	**0.456140**	**0.223180**
#05 - T. Camadas Cong.	Oboe	0.000000	0.000000	0.000000	0.004930	0.005780	0.004290	**0.016390**	**0.022560**	**0.012880**
#05 - T. Camadas Cong.	Master	0.221660	0.270000	0.190000	**0.278550**	**0.396830**	**0.214590**	0.262110	0.389830	0.197420

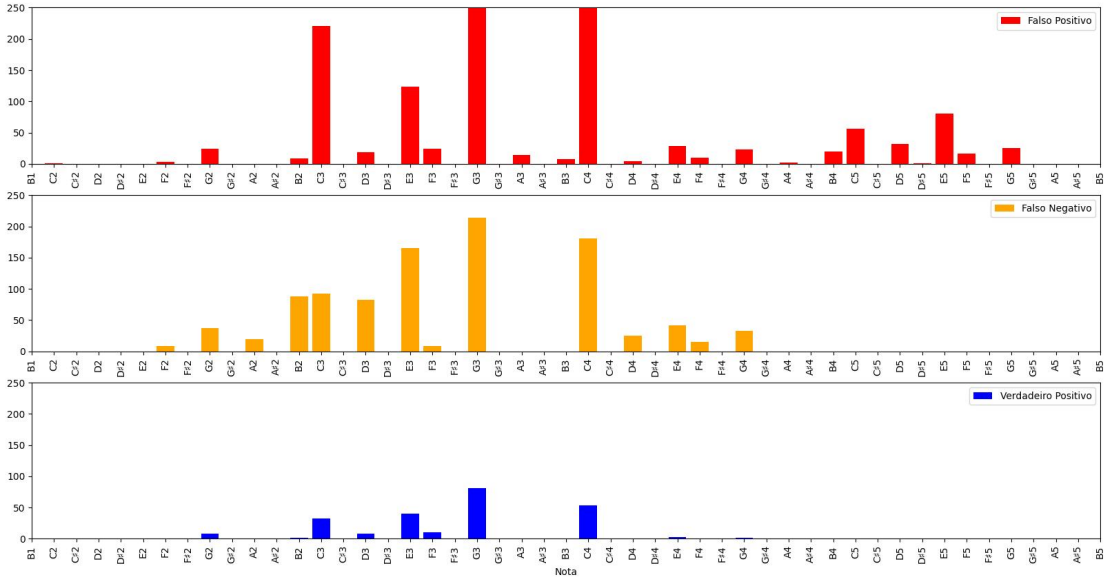
Tabela 6.17: Resultados O relógio bateu três horas - Variação dos registros. Configuração: Teclado com diferentes Timbres + Baixos

Figura 6.37: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #02 - 9900



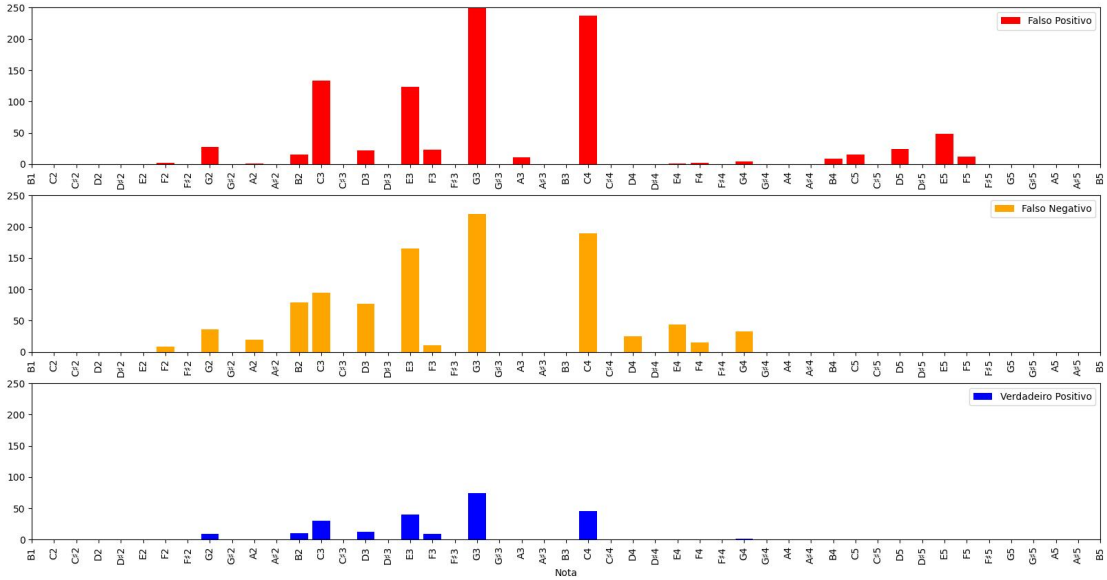
Fonte: Autoria Própria

Figura 6.38: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #03 - 3300



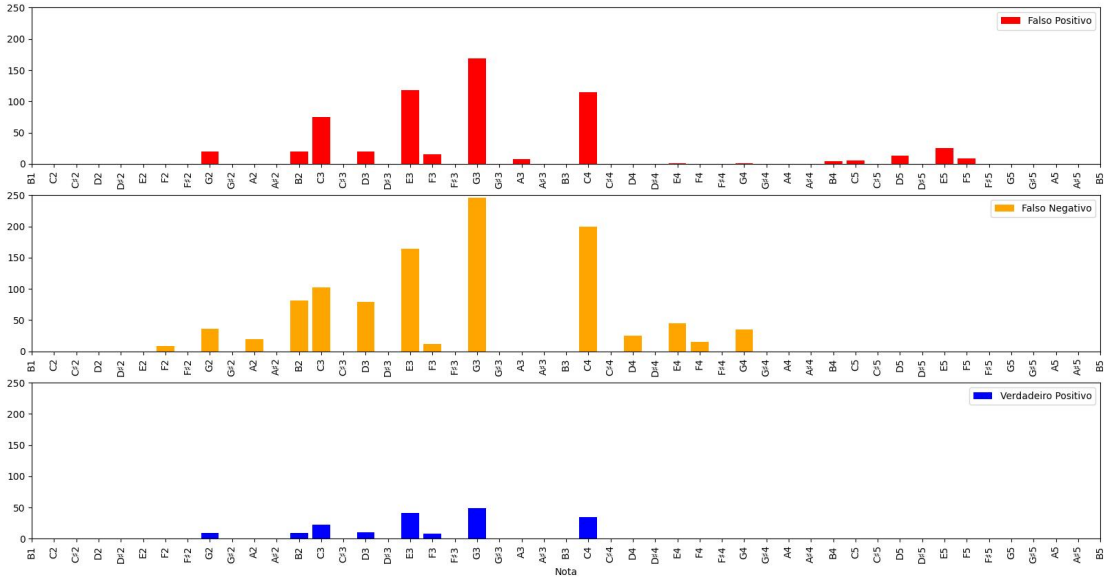
Fonte: Autoria Própria

Figura 6.39: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #03 - 6600



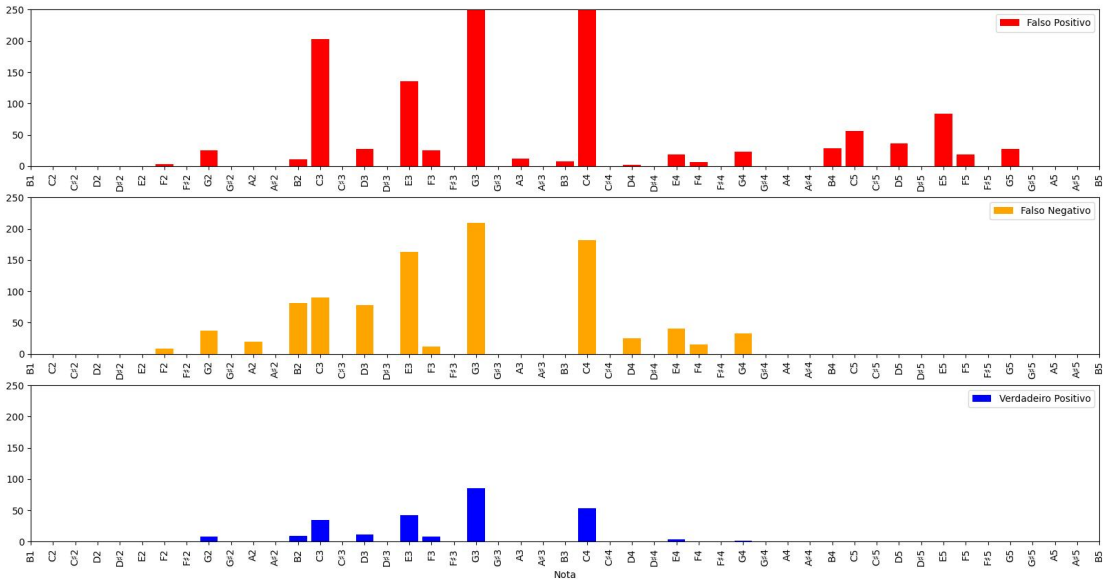
Fonte: Autoria Própria

Figura 6.40: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #03 - 9900



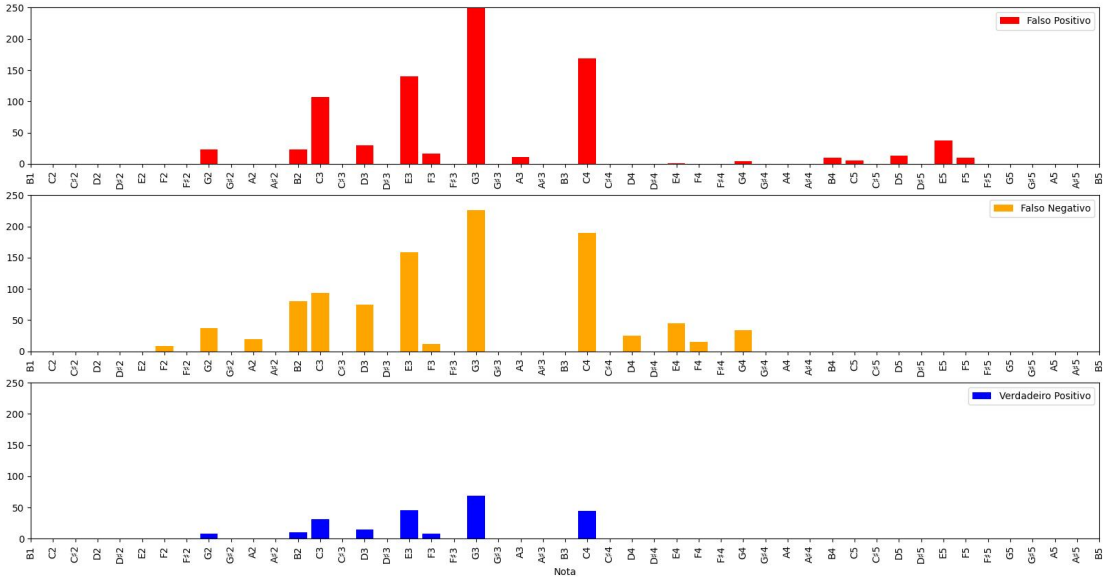
Fonte: Autoria Própria

Figura 6.41: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #04 - 3300



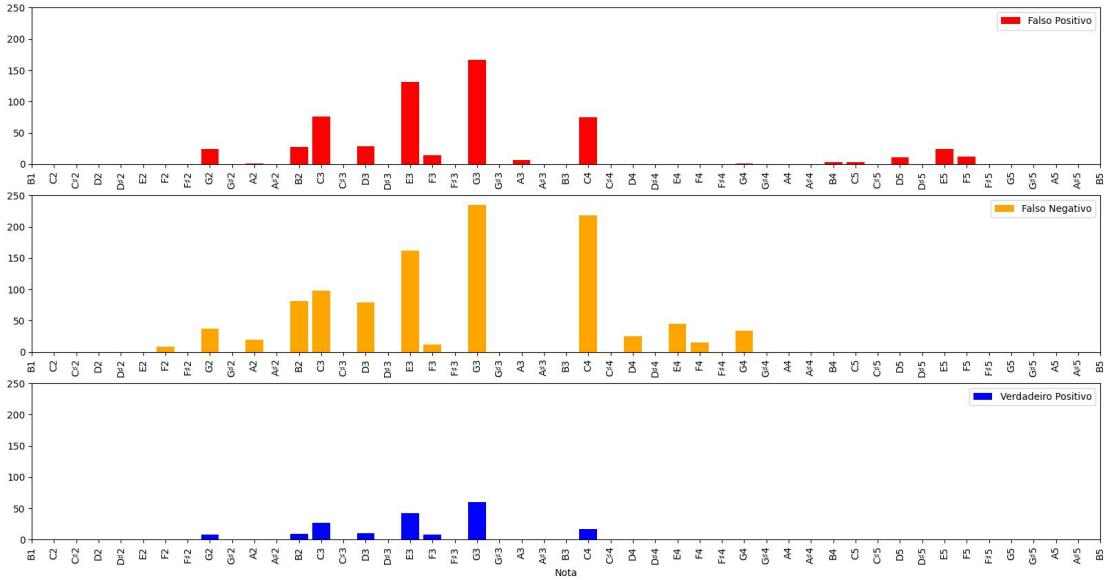
Fonte: Autoria Própria

Figura 6.42: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #04 - 6600



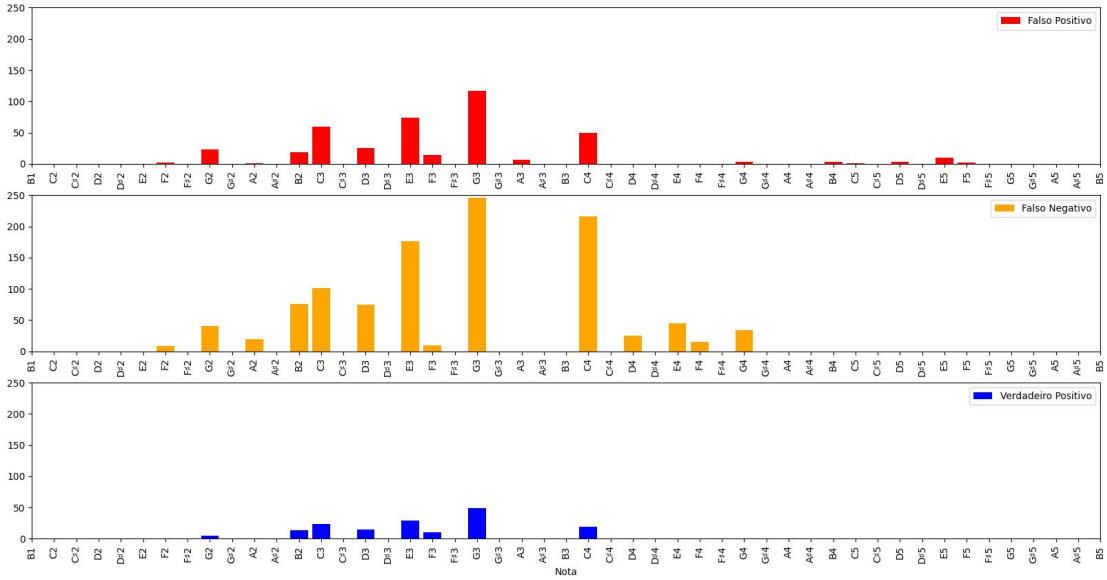
Fonte: Autoria Própria

Figura 6.43: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #04 - 9900



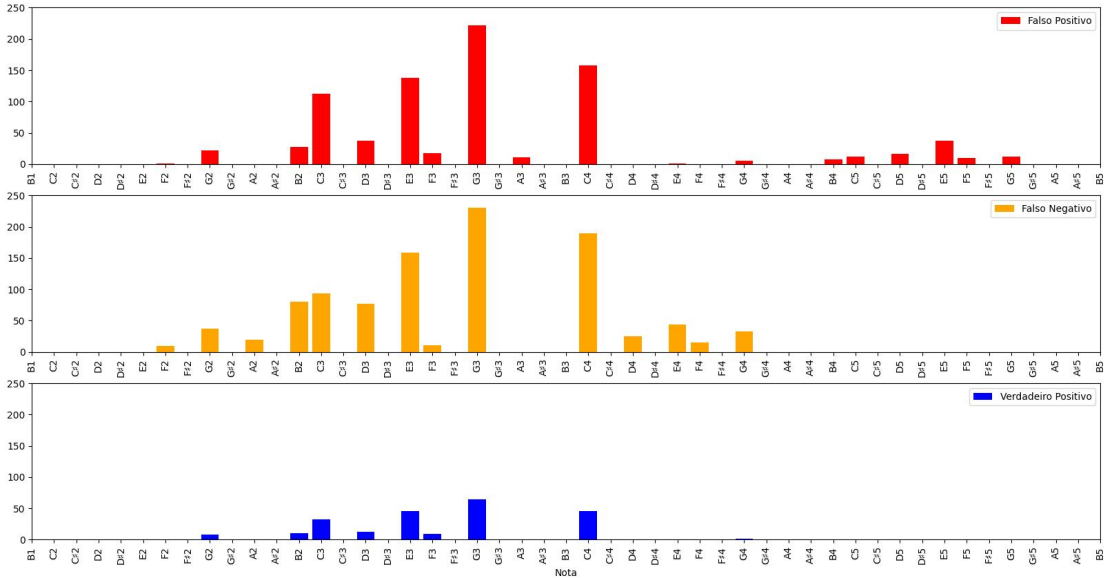
Fonte: Autoria Própria

Figura 6.44: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #05 - 3300



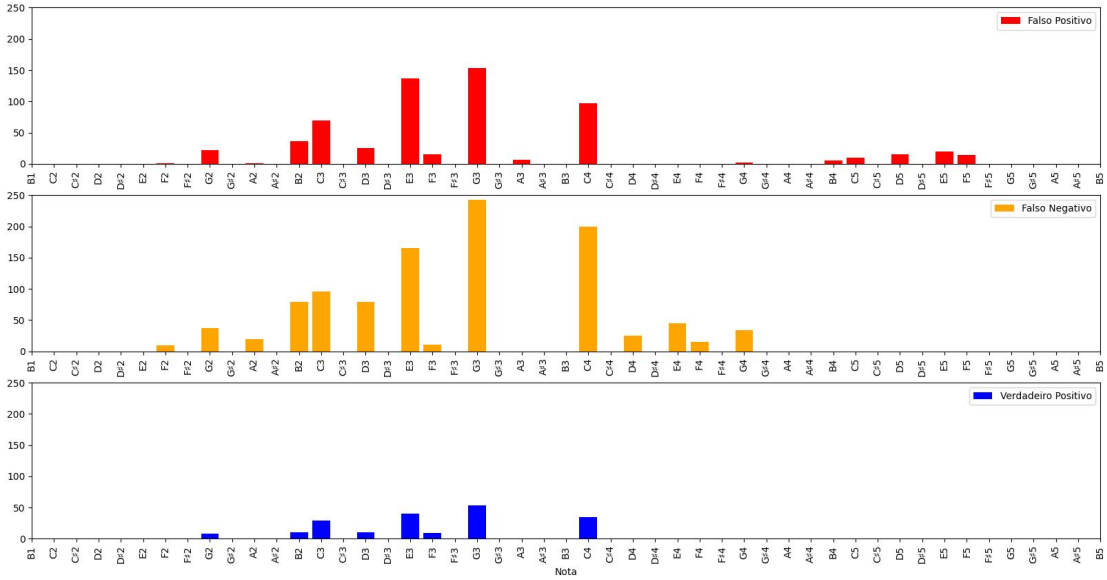
Fonte: Autoria Própria

Figura 6.45: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #05 - 6600



Fonte: Autoria Própria

Figura 6.46: Histograma Primeira Valsa (Teclado + Baixos) - Experimento #05 - 9900



Fonte: Autoria Própria

7

Conclusões

Este trabalho aborda a aplicação de técnicas de transferência de aprendizado em uma rede neural densa pré-treinada para o piano e sua capacidade de se adaptar a um instrumento diferente, porém similar, a Sanfona Stradella.

Ao decorrer do trabalho foi apresentado que apesar de ser um instrumento semelhante, a Sanfona tem suas peculiaridades e desafios específicos.

O primeiro desafio encontrado neste trabalho é que apesar da Sanfona ser um instrumento muito utilizado na Europa, no sul e nordeste do Brasil, Uruguai, Estados Unidos entre outros. Não há uma base aberta com gravações e suas respectivas transcrições alinhadas.

Após a construção da base, foram conduzidos seis experimentos cujo objetivo era responder as seguintes perguntas de pesquisa: (1) É possível realizar transferência de aprendizado com um classificador originalmente treinado para o piano para um novo classificador para a Sanfona? (2) Congelar diferentes números de camadas convolucionais altera a qualidade da transcrição? E qual o impacto no tempo de treinamento?; (3) Qual o impacto de se treinar a rede por mais épocas?; (4) Os diferentes registros da Sanfona impactam diretamente na transcrição?; (5) Qual o impacto do treinamento utilizando a base de Sanfona no problema original?; (6) Qual o impacto que os baixos da Sanfona na transcrição?

Após a execução dos experimentos, foi possível concluir que há um forte indício que é possível realizar a transferência de aprendizado do piano para Sanfona. Embora congelar diferentes números de camadas convolucionais da rede não ter demonstrado um impacto direto na qualidade da transcrição para músicas que utilizam tanto o baixo quanto o teclado da Sanfona, apresentando

uma diferença média menor que 0,05% para treinamentos acima de 6600 épocas.

O experimento #02 fine-tunning é o experimento que demonstrou ter o melhor desempenho como demonstrado na imagem 7.1. Já os experimentos #03, #04 e #05 demonstraram ter a mesma tendência como demonstrado pelas figuras 7.1, 7.2 e 7.3 indicando não haver uma relação direta entre o número de camadas convolucionais congeladas e a qualidade geral da transcrição, a não ser por uma diferença de recall no experimento #05 com 3300 épocas, diferença essa que se tornou imperceptível após 6600 épocas, como demonstrado na figura 7.3.

Congelar diferentes números de camadas convolucionais não afetou consideravelmente o tempo de treinamento, levando 1 hora e 30 minutos em média a cada 3300 épocas em um computador com 32gb de RAM, SSD e com uma placa de vídeo NVIDIA RTX 4080.

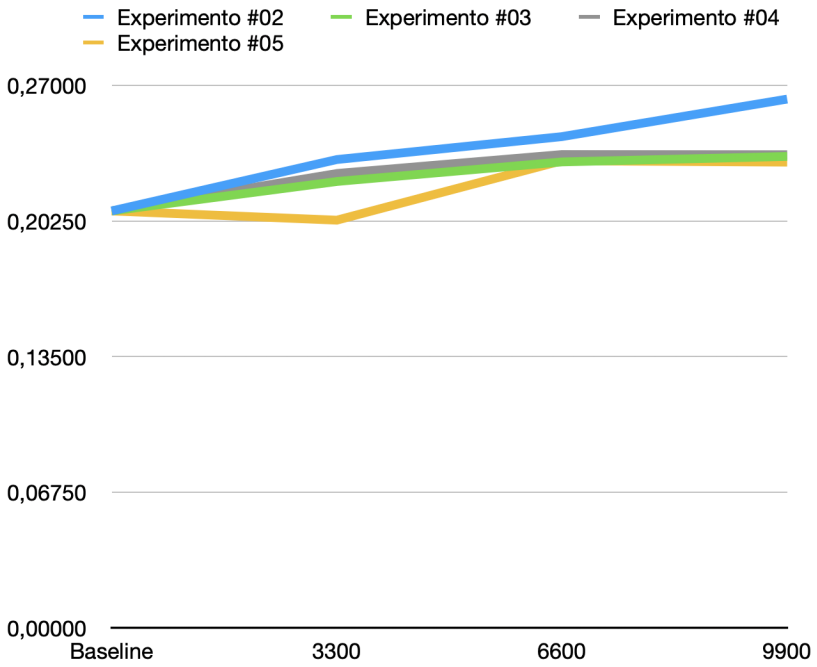
Os experimentos demonstraram um aumento na qualidade geral da transcrição (f1-score) conforme mais épocas eram adicionadas ao treinamento, onde com exceção do experimento #02 não demonstrando melhora entre 6600 e 9900 épocas de treinamento, isto por que ao mesmo tempo que o precision aumenta o recall continua diminuindo. O experimento #02 em 9900 épocas rompeu a tendência de aumento de precision e diminuição do recall demonstrando um ganho médio em ambas as métricas, esta ruptura pode ser um indício de underfitting.

Entretanto, outra possível explicação para este comportamento, é que as camadas convolucionais atuam como extratoras de características (LECUN et al., 1998), já que o experimento #02 não congela nenhuma camada convolucional. Isto sugere que diferentes instrumentos por produzirem diferentes timbres mudam significativamente o espaço de características do problema, interferindo, na qualidade do transfer-learning.

Ressaltando que o experimento #07 ao analisar músicas que contém apenas o teclado (lado direito da Sanfona) demonstraram ter uma queda na qualidade da transcrição.

Devido à baixa qualidade dos resultados apresentados no experimento #07 não é possível concluir se há um impacto direto na qualidade de transcrição causado pela variação dos registros do teclado, pois o experimento demonstrou a incapacidade do modelo de transcrever músicas que contenham apenas a melodia (teclado), sem a base rítmica (baixos). Por outro lado, o experimento #08 demonstrou que ao incluir os baixos, existe um grande impacto na qualidade da transcrição, tendo ambas as músicas uma diferença superior há 30% entre o registro Piccolo e Master com a configuração do experimento #02 e uma tendência

Figura 7.1: Tendência f1



Fonte: Autoria Própria

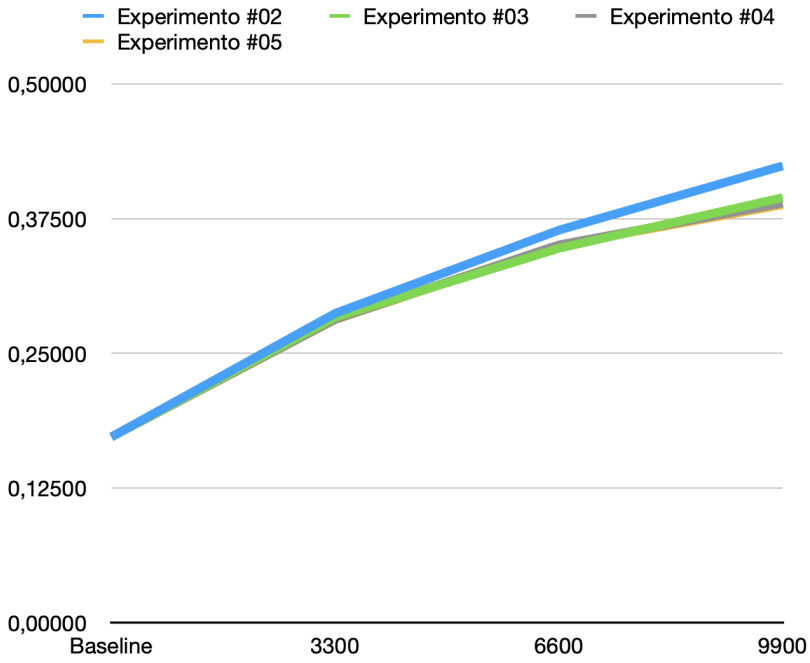
parecida ter sido apresentado para as demais configurações.

A perca de conhecimento do problema original é demonstrado pelo experimento #05, onde a tabela 6.10 demonstra as métricas utilizando os pesos originais da rede e a tabela 6.11 demonstra as métricas do piano conforme os demais experimentos avançam. Houve uma perca de média de precision de $\approx$ 4%, enquanto recall houve uma perca de $\approx$ 52%, ou seja, observa-se uma coerência com o padrão observado na Sanfona, caracterizado pela predominância de falsos negativos, sugerindo uma tendência do modelo a ignorar notas, enquanto as notas previstas pelo modelo exibem uma alta probabilidade de estarem corretas, ou seja, uma baixa tendência de falsos positivos.

7.1 Limitações e trabalhos futuros

A grande limitação deste trabalho é o tamanho de sua base Fole Dataset e a complexidade de suas músicas, contendo apenas aproximadamente meia hora de gravação divididas entre apenas 15 músicas, sendo que apenas duas

Figura 7.2: Tendência Precision



Fonte: Autoria Própria

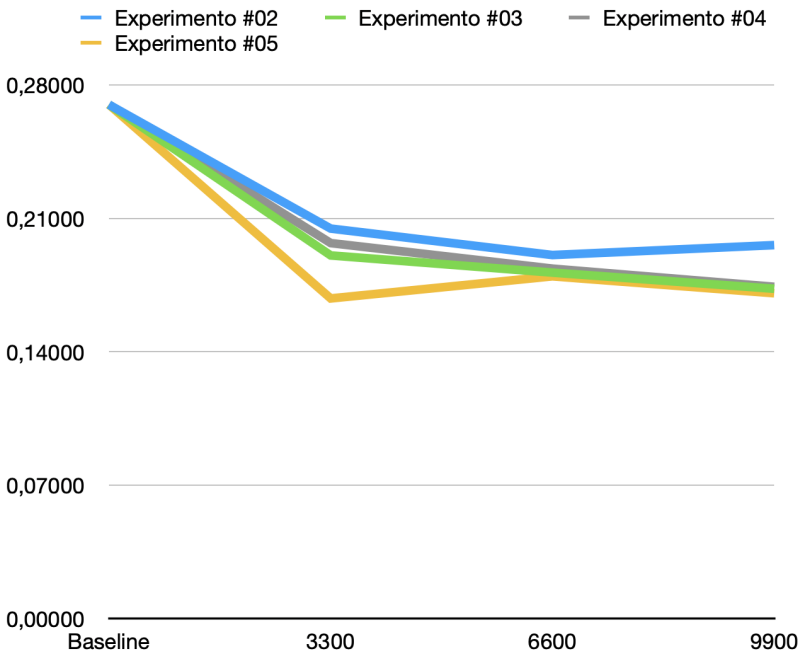
delas possuem variações de timbres e ambas as músicas de nível de dificuldade iniciante.

Outra limitação apresentada no Fole Dataset é a inclusão de apenas músicas ocidentais.

Como trabalho futuro pretende-se aumentar a base utilizando uma Sanfona eletrônico com saída midi para reduzir o trabalho de laboratório e aumentar a base exponencialmente, também é possível aplicar, e mesclar, diferentes técnicas de transferência de aprendizado como a transferência de aprendizado por instâncias e utilizar redes pré-treinadas de outros instrumentos como a flauta doce que seu timbre produz uma curva senoidal quase perfeita, por exemplo.

Uma abordagem diferente que pode ser explorada para realizar a transcrição de músicas de Sanfona é uso de mecanismos de atenção, em que (LU; ZHANG; NAYAK, 2020) em seu trabalhodemonstrou ter excelentes resultados para a classificação de áudios e o trabalho realizado por (WU; CHEN; SU, 2020) demonstra ótimos resultados para realizar a transcrição de músicas em músicas com diversos instrumentos.

Figura 7.3: Tendência Recall



Fonte: Autoria Própria

Referências

BAHETI, P. *What is transfer learning? [examples amp; newbie-friendly guide]*. 2021. Disponível em: <https://www.v7labs.com/blog/transfer-learning-guide>.

BAY, M.; EHMANN, A. F.; DOWNIE, J. S. Evaluation of multiple-f0 estimation and tracking systems. *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*, p. 315–320, 2009.

BENETOS, E.; DIXON, S.; DUAN, Z.; EWERT, S. Automatic Music Transcription: An Overview. *IEEE Signal Processing Magazine*, v. 36, n. 1, p. 20–30, jan. 2019. ISSN 1558-0792.

BöCK, S.; SCHEDL, M. Polyphonic piano note transcription with recurrent neural networks. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2012. p. 121–124. ISSN 2379-190X. ISSN: 2379-190X.

CAMARGO, M. *ACORDEOM BRASILEIRO*. [S.l.]: Independente, 2018.

CAMBRIDGE-WEBSTER. Sound. In: *Cambridge Dictionary*. [s.n.], 2023. Acessado em 17 de Jul. 2023. Disponível em: <https://dictionary.cambridge.org/dictionary/english/sound>.

CANNAM, C.; LANDONE, C.; SANDLER, M. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In: *Proceedings of the ACM Multimedia 2010 International Conference*. Firenze, Italy: [s.n.], 2010. p. 1467–1468.

CARTER, N. *Music Theory: From Beginner to Expert - The Ultimate Step-By-Step Guide to Understanding and Learning Music Theory Effortlessly*. CreateSpace Independent Publishing Platform, 2018. (Essential Learning Tools for Musicians). ISBN 9781986061834. Disponível em: <https://books.google.com.br/books?id=louZtAEACAAJ>.

ECHEVERRI, C. O.; RODRIGUEZ, J.; GARDUñO-APARICIO, M. An approach to stft and cwt learning through music hands-on labs. *Computer Applications in Engineering Education*, v. 26, 04 2018.

EMIYA, V.; BADEAU, R.; DAVID, B. Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 18, n. 6, p. 1643–1654, ago. 2010. ISSN 1558-7924.

EMIYA, V.; BERTIN, N.; DAVID, B.; BADEAU, R. *MAPS - A piano database for multipitch estimation and automatic transcription of music*. [S.l.], 2010. 11 p. Disponível em: <https://hal.inria.fr/inria-00544155>.

GÉRON, A. *Mãos à obra: Aprendizagem de máquina com Sckit-Learn e Tensorflow*. 1. ed. Rio de Janeiro: Alta Books Editora, 2019. Acesso em: 21 ago 2013.

GHOLAMIANGONABADI, D.; KISELOV, N.; GROLINGER, K. Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection. *IEEE Access*, Institute of Electrical and Electronics Engineers (IEEE), v. 8, p. 133982–133994, 2020. ISSN 2169-3536. Disponível em: <http://dx.doi.org/10.1109/ACCESS.2020.3010715>.

GOTO, M. Física e música em consonância. *Revista Brasileira de Ensino de Física*, FapUNIFESP (SciELO), v. 31, n. 2, p. 2307.1–2307.8, jun. 2009. Disponível em: <https://doi.org/10.1590/s1806-11172009000200008>.

HALLIDAY, D.; RESNICK, R.; WALKER, J. *Fundamentos de física: volume 2 : gravitação, ondas e termodinâmica*. [S.l.: s.n.], 2008.

HAWTHORNE, C.; ELSEN, E.; SONG, J.; ROBERTS, A.; SIMON, I.; RAFFEL, C.; ENGEL, J.; OORE, S.; ECK, D. Onsets and Frames: Dual-Objective Piano Transcription. *arXiv:1710.11153 [cs, eess, stat]*, jun. 2018. ArXiv: 1710.11153. Disponível em: <http://arxiv.org/abs/1710.11153>.

HAWTHORNE, C.; STASYUK, A.; ROBERTS, A.; SIMON, I.; HUANG, C.-Z. A.; DIELEMAN, S.; ELSEN, E.; ENGEL, J.; ECK, D. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. *arXiv:1810.12247 [cs, eess, stat]*, jan. 2019. ArXiv: 1810.12247. Disponível em: <http://arxiv.org/abs/1810.12247>.

HOSSAIN, I.; KHOSRAVI, A.; HETTIARACHCHI, I.; NAHAVANDI, S. Multiclass informative instance transfer learning framework for motor imagery-based brain-computer interface. *Computational Intelligence and Neuroscience*, Hindawi Limited, v. 2018, p. 1–12, 2018. Disponível em: <https://doi.org/10.1155/2018/6323414>.

KELZ, R.; DORFER, M.; KORZENIOWSKI, F.; BöCK, S.; ARZT, A.; WIDMER, G. On the Potential of Simple Framewise Approaches to Piano Transcription. *arXiv:1612.05153 [cs]*, dez. 2016. ArXiv: 1612.05153. Disponível em: <http://arxiv.org/abs/1612.05153>.

KLEBER, M. Distintos ventos dos foles: dos primeiros fonogramas ao modismo do acordeão na década de 1950 no brasil. *Simpósio Brasileiro de Pós Graduação em Música*, v. 5, 06 2018.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998.

LU, H.; ZHANG, H.; NAYAK, A. *A Deep Neural Network for Audio Classification with a Classifier Attention Mechanism*. 2020. Disponível em: <https://arxiv.org/abs/2006.09815>.

MACIEJEWSKI, T.; LUKASIK, E. Accordion Music and its Automatic Transcription to MIDI Format. In: *Audio Engineering Society Convention 134*. [s.n.], 2013. Disponível em: <http://www.aes.org/e-lib/browse.cfm?elib=16728>.

MALLAWAARACHCHI, V. *Label propagation demystified*. Towards Data Science, 2020. Disponível em: <https://towardsdatascience.com/label-propagation-demystified-cd5390f27472>.

MASCARENHAS, M. *Metódo de acordeão Mascarenhas*. [S.l.]: Ricordi Brasileira, 2006. ISBN 8599477269.

MAZZOLA, G.; PANG, Y.; HEINZE, W.; GKOUDINA, K.; PUJAKUSUMA, G. A.; GRUNKLEE, J.; CHEN, Z.; HU, T.; MA, Y. A short history of midi. In: ____. *Basic Music Technology: An Introduction*. Cham: Springer International Publishing, 2018. p. 115–116. ISBN 978-3-030-00982-3. Disponível em: <https://doi.org/10.1007/978-3-030-00982-3_11>.

MCLEOD, A.; STEEDMAN, M. Evaluating automatic polyphonic music transcription. In: *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*. [S.l.]: International Society for Music Information Retrieval, 2018. p. 42–49. ISBN 9782954035123.

MED, B. *Teoria da música 4 ed.* [S.l.]: MusiMed, 1996.

MIDO, L. *MIDI files¶*. 2021. Disponível em: <https://mido.readthedocs.io/en/latest/midi_files.html>.

MONIGATTI, L. *Audio classification with Deep Learning in python*. Towards Data Science, 2023. Disponível em: <https://towardsdatascience.com/audio-classification-with-deep-learning-in-python-cf752b22ba07>.

PAIVA, C. N. d. Uma experiência de ensino do acordeon na escola de música da UFRN. Universidade Federal do Rio Grande do Norte, 2014. Disponível em: <http://monografias.ufrn.br/handle/123456789/1389>.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 22, n. 10, p. 1345–1359, 2010.

PEDERSEN, P. The mel scale. *Journal of Music Theory*, Duke University Press, v. 9, n. 2, p. 295, 1965. Disponível em: <https://doi.org/10.2307/843164>.

SHATRI, E.; FAZEKAS, G. Optical music recognition: State of the art and major challenges. *CoRR*, abs/2006.07885, 2020. Disponível em: <https://arxiv.org/abs/2006.07885>.

SIGTIA, S.; BENETOS, E.; DIXON, S. An End-to-End Neural Network for Polyphonic Piano Music Transcription. *arXiv:1508.01774 [cs, stat]*, fev. 2016. ArXiv: 1508.01774. Disponível em: <http://arxiv.org/abs/1508.01774>.

SILVA, D. C. M. d. *Timbre. O que É timbre?* Mundo Educação, 2023. Disponível em: <https://mundoeducacao.uol.com.br/fisica/timbre.htm>.

STANFORD-WEBMASTER. *Mel — sfu.ca*. s.d. <https://www.sfu.ca/sonic-studio-webdav/handbook/Mel.html>. [Accessed 30-07-2023].

SU, L.; YANG, Y.-H. Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In: KRONLAND-MARTINET, R.; ARAMAKI, M.; YSTAD, S. (Ed.). *Music, Mind, and Embodiment*. Cham: Springer International Publishing, 2016. p. 309–321. ISBN 978-3-319-46282-0.

VILALTA, R.; GIRAUD-CARRIER, C.; BRAZDIL, P.; SOARES, C. Inductive transfer. In: ____. *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2010. p. 545–548. ISBN 978-0-387-30164-8. Disponível em: <https://doi.org/10.1007/978-0-387-30164-8_401>.

WEI, Y.; ZHANG, Y.; HUANG, J.; YANG, Q. Transfer learning via learning to transfer. In: DY, J.; KRAUSE, A. (Ed.). *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018. (Proceedings of Machine Learning Research, v. 80), p. 5085–5094. Disponível em: <https://proceedings.mlr.press/v80/wei18a.html>.

WIKIPEDIA-WEBSTER. Sheng (instrument). In: *Sheng (instrument)*. [s.n.], 2024. Acessado em 28 de Fev. 2024. Disponível em: <https://en.wikipedia.org/wiki/Sheng_%28instrument%29>.

WU, Y.-T.; CHEN, B.; SU, L. Multi-instrument automatic music transcription with self-attention-based instance segmentation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 28, p. 2796–2809, 2020.

ZHUANG, F.; QI, Z.; DUAN, K.; XI, D.; ZHU, Y.; ZHU, H.; XIONG, H.; HE, Q. *A Comprehensive Survey on Transfer Learning*. 2020.