TRABALHO DE GRADUAÇÃO

UM ESTUDO SOBRE A CLASSIFICAÇÃO E SEPARAÇÃO ON-LINE DE FONTE SONORA MONOFÔNICA USANDO REDES NEURAIS RECORRENTES RASAS E PROFUNDAS

Daniel Bauchspiess

Brasília, Maio de 2021

ENGENHARIA
MECATRÔNICA
UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia
Curso de Graduação em Engenharia de Controle e Automação

TRABALHO DE GRADUAÇÃO

UM ESTUDO SOBRE A CLASSIFICAÇÃO E SEPARAÇÃO ON-LINE DE FONTE SONORA MONOFÔNICA USANDO REDES NEURAIS RECORRENTES RASAS E PROFUNDAS

Daniel Bauchspiess

Relatório submetido como requisito parcial de obtenção de grau de Engenheiro de Controle e Automação

Banca Examinadora

Prof. Dr. Marcus Vinicius Lamar, CIC/UnB
Orientador

Prof. Dra. Tatiana O. Catanzaro, MUS/UnB
Coorientadora

Prof. Dr. Márcio Brandão, CIC/UnB
Membro Interno - Aposentado

Profa. Dra. Carla Maria C. C. Koike, CIC/UnB
Membro Interno

Brasília, Maio de 2021

FICHA CATALOGRÁFICA

BAUCHSPIESS, DANIEL

Um estudo sobre a classificação e separação on-line de fonte sonora monofônica usando redes neurais recorrentes rasas e profundas,

[Distrito Federal] 2021.

viii, 60p., 297 mm (FT/UnB, Engenheiro, Controle e Automação, 2021). Trabalho de Graduação – Universidade de Brasília.Faculdade de Tecnologia.

1. Classificação de Fonte Sonora 2.Separação de Fonte Sonora
3. Redes Neurais

I. Mecatrônica/FT/UnB II. Título (Série)

REFERÊNCIA BIBLIOGRÁFICA

BAUCHSPIESS, D., (2021). Um estudo sobre a classificação e separação on-line de fonte sonora monofônica usando redes neurais recorrentes rasas e profundas. Trabalho de Graduação em Engenharia de Controle e Automação, Publicação FT.TG-$n°$04, Faculdade de Tecnologia, Universidade de Brasília, Brasília, DF, 60p.

CESSÃO DE DIREITOS

AUTOR: Daniel Bauchspiess

TÍTULO DO TRABALHO DE GRADUAÇÃO: Um estudo sobre a classificação e separação on-line de fonte sonora monofônica usando redes neurais recorrentes rasas e profundas.

GRAU: Engenheiro ANO: 2021

Daniel Bauchspiess

SQN 208, Asa Norte.

70853-060 Brasília – DF – Brasil.

Dedicatória

Dedico este trabalho a todos aqueles que fizeram parte desta jornada, tanto direta quanto indiretamente, seja com agregação de conteúdo ou com um simples mas importantíssimo apoio moral e emocional.

Daniel Bauchspiess

Agradecimentos

Gostaria de agradecer primeiramente à minha família, por sempre me dar muito apoio tanto nos estudos quanto emocionalmente, e que nesse período de pandemia foram minha rocha. Uma nota de reconhecimento vai para minha irmã, Cristina, responsável pela adaptação das figuras utilizadas para representar características de redes neurais.
Agradeço aos meus orientadores, Marcus e Tatiana, pelo período passado no desenvolvimento deste trabalho de graduação, por toda a paciência e disposição demonstrada. Foi uma experiência muito enriquecedora.

Daniel Bauchspiess

RESUMO

Neste trabalho, são estudados o uso de redes neurais para abordar dois problemas: classificação de instrumentos musicais e separação de fonte sonora monofônica. Para a classificação, as redes estudadas foram das arquiteturas Multilayer Perceptron e recorrente de Elman, recebendo como sinal de entrada cinco descritores sonoros correspondentes a um trecho de um sinal de áudio de um instrumento musical ao longo do tempo. A classificação é feita em cada trecho do sinal de áudio, devendo classificá-lo como som de contrabaixo ou flauta. O melhor resultado para treinamento em apenas uma nota de cada instrumento alcançou uma acurácia de 97,19%, e o melhor resultado quando treinado em 6 notas de cada instrumento alcançou 96,44% de acurácia. Para a tarefa de separação on-line de uma fonte sonora mixada, em seus canais componentes, a partir da análise sequencial de suas amostras no tempo, são analisadas redes recorrentes rasas de Elman e profundas LSTM. Os resultados obtidos indicam que as estruturas neurais testadas são promissoras e podem atingir alta qualidade na separação, tendo o SDR como figura de mérito.

Palavras Chave: Redes neurais, Multilayer Perceptron, rede de Elman, LSTM, Música, Espectro de frequências, Descritores sonoros, Classificação de fonte sonora, Separação de fonte sonora

ABSTRACT

In this project, neural networks were developed to tackle two problems: musical instrument classification and monophonic sound source separation. For the classification phase, the developed networks were the Multilayer Perceptron and the recurrent Elman, being fed five sound descriptors related to a chunk of a musical instrument audio signal as input. Each audio signal chunk is classified between contrabass and flute. The best result from training in one note of each instrument achieved 97.19% accuracy, and the best training in six notes of each instrument achieved 96.44% accuracy. In the subject of online separation of a mixed sound source, Elman shallow neural networks and deep LSTM are inspected through sequential analysis of the samples in its channels in time. The results show that the tested neural structures are promising and can give a high quality separation using SDR as evaluation metrics.

Keywords: Neural networks, Multilayer Perceptron, Elman network, LSTM, Music, Frequency spectrum, Sound descriptors, Sound source classification, Sound source separation

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

Símbolos Latinos

x_i	Sinal de entrada de neurônio ou de rede neural
w_i	Pesos associados às entradas dos neurônios
u	Potencial de ativação do neurônio
$g(.)$	Função de ativação do neurônio
y	Sinal de saída de um neurônio
f_t	Vetor de ativação da porta de esquecimento
i_t	Vetor de ativação da porta de escrita
o_t	Vetor de ativação da porta de saída
h_t	Vetor de saída da unidade LSTM
$\tilde{c_t}$	Vetor de ativação da célula de memória
c_t	Vetor de estado da célula de memória
W, U	Matriz de pesos de uma rede
b	Vetor de bias
h	Número de características de entrada da célula de memória
d	Número de unidades escondidas de células de memória
n	Metade do tamanho de uma janela espectral
$a[i]$	Amplitude do sinal na frequência indicada pelo índice i
$slope/decrease$	Declive/Decrescimento Espectral
X_d	Vetor de Descritores
$\hat{s}$	Sinal estimado
s_{target}	Sinal esperado modificado por distorção permitida
e_{interf}	Interferências vindas de outras fontes sonoras
e_{noise}	Ruído presente no sinal
e_{artif}	Artefatos presentes no sinal
L	Limiar de amplitude
A	Amplitudes absolutas médias de janelas de um sinal
$m(t)$	Sinal composto pela mistura de fontes no tempo t
N_F	Número de fontes presentes na mistura
$f_i(t)$	Sinal da fonte i no tempo t
$f_{ei}(t)$	Fontes estimadas
S	Número de amostras em um sinal

Símbolos Gregos

$\sum$	Agregador linear
θ	Bias
σ_g	Função de ativação sigmoidal
σ_c	Função de ativação tangente hiperbólico
μ	Centroide Espectral
ν	Espalhamento Espectral

Grupos Adimensionais

K	Quantidade de grupos para treino com K-Fold

Siglas

MIR	*Music Information Retrieval*
RNA	*Rede Neural Artificial*
MLP	*Multilayer Perceptron*
LSTM	*Long Short-Term Memory*
FT	*Fourier Transform*
STFT	*Short-Time Fourier Transform*
SDR	*Source to Distortion Ratio*
SIR	*Source to Interferences Ratio*
SNR	*Source to Noise Ratio*
SAR	*Source to Artifacts Ratio*

Capítulo 1

Introdução

1.1 Contextualização

Dentro de uma peça de música, existem muitas informações que podem ser extraídas e que seres humanos reconhecem de forma natural. Informações mais gerais como o gênero musical ou os instrumentos tocados, ou informações mais locais como a força, o ritmo da música, dentre outras. A capacidade de extrair essas informações de forma automática pode abrir muitas portas na área acadêmica e comercial, possibilitando estudos aprofundados de cada peça, seleção de músicas semelhantes, sistemas que respondem ao andamento da música e assim por diante. A área crescente que estuda as formas de extrair essas informações é a de Recuperação da Informação da Música (MIR, do inglês *Music Information Retrieval*) [5].

1.2 Definição do problema

Dentre os ramos de pesquisa de MIR, existe a área de identificação e separação de fonte sonora. Um problema que costuma ser citado nessas áreas é o da "festa de cocktail", abordada por Cherry [6]: em uma sala cheia de pessoas conversando, o ser humano é capaz de focar em uma conversa, ignorando o ruído do ambiente. Para o caso de sinais musicais o problema da separação é diferente, não havendo uma única fonte de interesse a ser diferenciada do ruído de fundo, mas uma variedade de tons e timbres tocando de forma coordenada [4].

Além de simplesmente diferenciar uma fonte de outra, o ser humano é capaz de identificá-la, caso seja conhecida. Em uma conversa, consegue reconhecer a voz da pessoa com quem está conversando. Ao escutar uma música, pode reconhecer a voz de algum cantor específico ou os instrumentos que são utilizados na peça.

Considerando essas situações, é trazido o problema da classificação e/ou separação de fontes sonoras de forma automática. Muitos sistemas já foram propostos nessas áreas [7][3][8][9][4] , mas não se encontram muitos estudos a respeito do uso de redes puramente recorrentes que realizem a classificação e separação *online*.

1.3 Objetivo Geral

Este projeto é voltado aos problemas da classificação e da separação de fontes sonoras musicais. Cada um desses problemas será abordado separadamente. O objetivo é comparar o desempenho de diferentes arquiteturas de redes neurais nos desafios propostos, em especial redes neurais recorrentes, e explorar as limitações de cada uma em seu contexto.

1.4 Objetivos específicos

Para a classificação de fonte sonora, o objetivo é comparar o desempenho de redes *feedforward* e recorrente na classificação de pequenos trechos de sons instrumentais baseado em descritores retirados de cada trecho. As redes criadas devem ser capazes de realizar a separação *online*.

Para a separação de fonte sonora, cada um dos sistemas propostos deve ser capaz de separar os sinais que compõem uma mistura a cada amostra apresentada. Neste contexto, o objetivo é realizar um estudo sobre arquiteturas de redes neurais recorrentes rasas e profundas, assim como alguns métodos para recuperar a rede com melhor separação do treinamento. Diferentes complexidades do problema serão abordadas, com sinais simples artificiais e de gravações instrumentais, assim como o treino especializado e o generalizado.

1.5 Apresentação do manuscrito

Este documento é composto de cinco capítulos. Este Capítulo 1 - Introdução - apresenta os objetivos do projeto, com os problemas a serem tratados e seu contexto. No Capítulo 2 - Fundamentos - será apresentado o embasamento teórico necessário para compreender os temas abordados pelo projeto, passando por assuntos de redes neurais e de acústica, e também serão apresentados alguns trabalhos recentes relacionados a este tema. No Capítulo 3 - Metodologia - será apresentada a metodologia empregada, o banco de dados utilizado e as redes sob análise, juntamente com suas estruturas e o contexto em que os experimentos foram feitos. O Capítulo 4 - Resultados - apresenta cada uma das configurações testadas dos sistemas e seus resultados, junto de uma análise crítica dos mesmos. Por fim, no Capítulo 5 - Conclusões - serão resumidas algumas das principais conclusões dos resultados obtidos em relação aos objetivos do projeto, seguido de proposições de trabalhos futuros.

Capítulo 2

Fundamentos

Para que se possa compreender as redes criadas e os estudos feitos no projeto, faz-se necessário fornecer uma base dos conteúdos abordados. Adiante serão tratados alguns temas de redes neurais e acústica, assim como formas utilizadas para avaliação quantitativa do desempenho de sistemas de classificação e separação de fonte sonora, junto com alguns trabalhos de outros autores realizados nessas áreas.

2.1 Redes Neurais Artificiais

Redes Neurais Artificiais (RNA) são representações computacionais e matemáticas inspiradas no sistema nervoso de seres vivos, cujo objetivo é mimetizar o processo de aprendizagem dos mesmos. Como é afirmado por Silva, I. Nunes, tradução livre, [10], página 5:

> "Redes neurais artificiais são modelos computacionais inspirados pelo sistema nervoso de seres vivos. Elas possuem a habilidade de adquirir e manter conhecimento (baseadas em informação) e podem ser definidas como um conjunto de unidades processadoras, representadas por neurônios artificiais, interligadas por muitas interconexões."

Como fora afirmado, as unidades processadoras de uma rede neural são os neurônios. O modelo mais simples de neurônio artificial, proposto por McCulloch e Pitts [11] em 1943 e apresentando as características principais de um neurônio biológico, pode ser implementado como na Figura 2.1 [10].

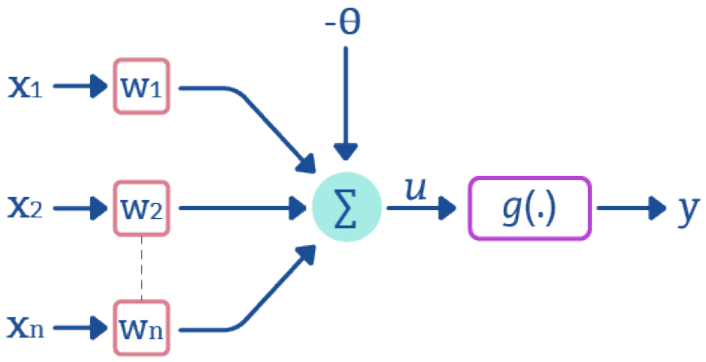


Figura 2.1: Implementação do modelo de neurônio proposto por McCulloch e Pitts

Nesse modelo, destaca-se os seguintes elementos:

- *Sinal de entrada do neurônio*, representado pelo vetor $\{x_1, x_2, ..., x_n\}$.
- *Pesos associados a cada entrada*, representados pelo vetor $\{w_1, w_2, ..., w_n\}$.
- *Agregador linear*, representado pelo símbolo do somatório $\sum$.
- *Bias*, representado pelo símbolo θ.
- *Potencial de ativação*, representado por u.
- *Função de ativação*, representado por $g(\cdot)$.
- *Sinal de saída*, representado por y.

Este sistema é regido pelas equações 2.1 e 2.2:

$$u = \sum_{i=1}^{n} w_i x_i - \theta, \text{ e} \tag{2.1}$$

$$y = g(u). \tag{2.2}$$

Por meio destas, pode-se interpretar u como uma soma ponderada das entradas (menos o *bias*). Originalmente, $g(u)$ fora proposto como uma Unidade de Lógica de Limiar [11][12], em que o resultado é 1 caso u esteja acima de certo valor, e 0 caso contrário. Atualmente outras funções de ativação também são aplicadas [12]. A saída do neurônio é dada por y, podendo esta ser aplicada como entrada de outros neurônios ou usada como saída da rede neural.

2.1.1 Redes multiple-layer feedforward

Uma rede neural é formada por neurônios conectados entre si e com a entrada, e estes são agrupados em camadas. Os três tipos de camadas são:

- *Camada de entrada*, responsável por receber o sinal de entrada externo.
- *Camadas intermediárias, escondidas ou invisíveis*, formada por neurônios responsáveis por extrair padrões associados ao processo sendo analisado.
- *Camada de saída*, também formada por neurônios, cujas saídas serão a saída da rede neural, resultado do processamento dos neurônios das camadas anteriores.

A Figura 2.2 mostra o exemplo de uma rede neural.

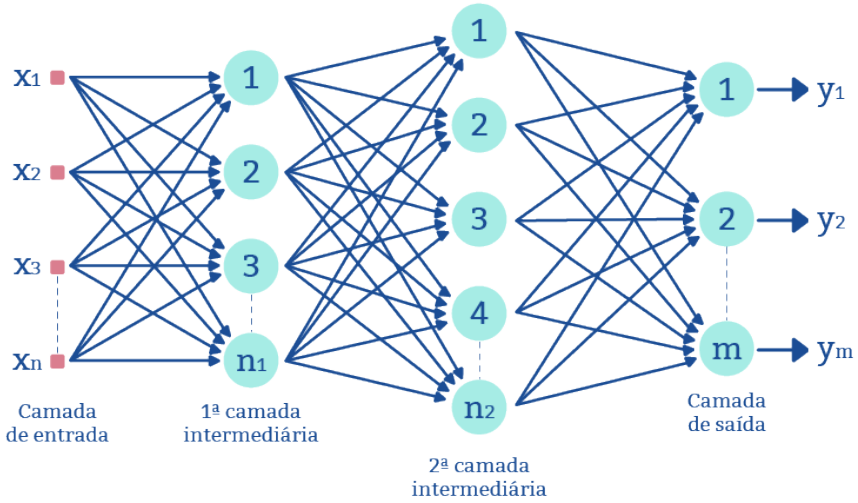


Figura 2.2: Representação de uma rede multiple-layer feedforward.

Ela é composta da camada de entrada, duas camadas intermediárias com n_1 e n_2 neurônios, e uma camada de saída com m neurônios. A disposição dos neurônios, como eles estão interconectados e como as camadas da rede são compostas caracterizam as principais arquiteturas de redes neurais [10].

A rede apresentada na Figura 2.2 é do tipo multiple-layer feedforward. Estas possuem uma ou mais camadas intermediárias (além das camadas de entrada e saída), e o fluxo da informação sempre segue uma mesma direção, da entrada até a saída [10]. Isto significa que as saídas dos neurônios de uma camada conectam-se apenas com os neurônios de camadas posteriores, mais próximas à camada de saída, nunca com a própria camada ou com camadas anteriores.

Dentre as redes que utilizam a arquitetura multiple-layer feedforward, uma das principais é a Multilayer Perceptron (MLP). Como mencionado para redes feedforward, esta rede possui pelo menos uma camada intermediária e o fluxo de informação é unidirecional, da entrada à saída da rede. Além disso, seu treinamento é supervisionado e a quantidade de neurônios em cada camada é variável.

2.1.2 Redes recorrentes

Redes recorrentes são caracterizadas pelo fato de que a saída de um ou mais neurônios serve de retroalimentação para outros neurônios [10], servindo de entrada de neurônios na mesma camada ou em camadas anteriores, fazendo com que o fluxo de informação não seja unidirecional. A Figura 2.3 apresenta um exemplo de uma rede neural recorrente.

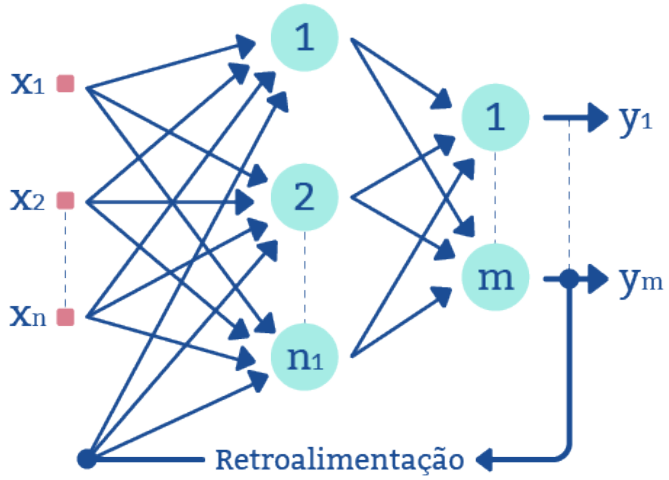


Figura 2.3: Representação de uma rede recorrente.

Nela, a saída do neurônio `m` da camada de saída é utilizada como entrada na camada intermediária. Essa retroalimentação pode ser associada a uma ou mais unidades de atraso. Isso significa que o sinal que está sendo passado não será utilizado como entrada dos neurônios em conjunto com o sinal de entrada que o gerou, e sim com o sinal de entradas posteriores. Por exemplo, se na rede da Figura 2.3 a retroalimentação possuir uma unidade de atraso, a saída do neurônio `m` do instante t será aplicada na camada intermediária apenas no instante $t + 1$, junto com as entradas desse mesmo instante.

O atributo de retroalimentação permite que essas redes sejam utilizadas em sistemas variantes no tempo [10], considerando que cada classificação considerará não só o dado inserido na rede em um determinado instante de tempo, mas também os dados anteriores a este. Um exemplo de rede desta arquitetura é a rede de Elman.

2.1.2.1 Rede de Elman

Proposta por Elman em 1990 [13], a rede que leva seu nome é um tipo específico de rede recorrente. Sua representação pode ser vista na Figura 2.4.

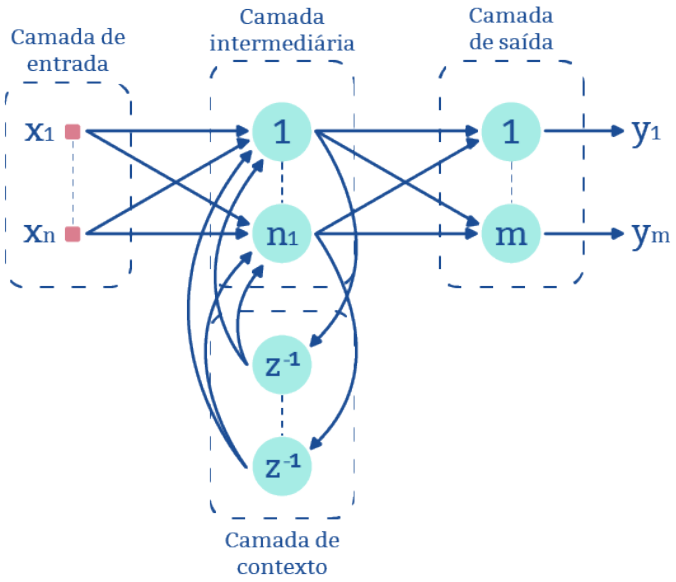


Figura 2.4: Representação de uma rede de Elman, adaptada de [1].

Ela é composta pela camada de entrada, uma camada intermediária e a camada de saída. Cada neurônio da camada intermediária é retroalimentada a todos os outros neurônios da mesma camada com uma unidade de atraso (representado pelas unidades z^{-1}), permitindo que a rede reconheça padrões temporais [14].

2.1.2.2 Long Short-Term Memory (LSTM)

Ao treinar uma rede neural, geralmente se utiliza um gradiente para determinar como os pesos da rede serão alterados, de forma a minimizar o erro (diferença entre as saídas obtidas e esperadas). Para redes recorrentes com sequências de entrada longas, entretanto, costuma-se encontrar o problema da explosão ou desaparecimento do gradiente: o gradiente terá um valor extremamente alto (explosão) ou extremamente baixo (desaparecimento), e quanto maior for a sequência, mais o problema se agrava. Quando o gradiente é muito baixo, o progresso do treinamento será muito pequeno, impedindo que a rede seja treinada em tempo hábil. Quando ele é muito alto, o treinamento torna-se instável [15].

Direcionados a este problema, Hochreiter e Schmidhuber propõem a Long Short-Term Memory (LSTM) [16], a qual ainda recebera atualizações com o tempo. A ideia por trás da arquitetura da LSTM é uma célula de memória capaz de manter seu estado ao longo do tempo, com portas que regulam a entrada e saída de informação da mesma [17]. A Figura 2.5 apresenta um modelo de uma célula de memória para a arquitetura LSTM.

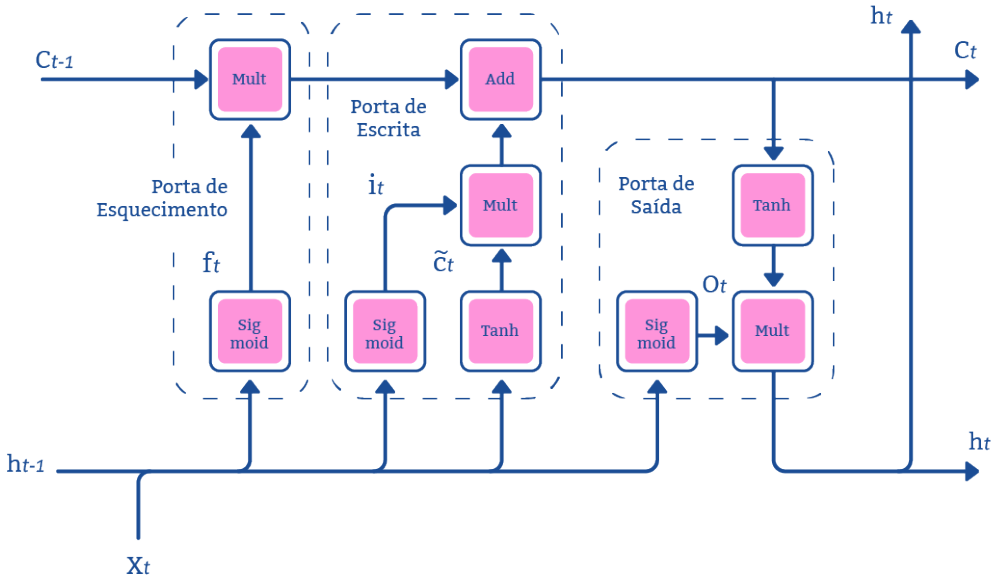


Figura 2.5: Representação de uma célula de memória utilizada na arquitetura LSTM. Adaptada de [2]

Nela, pode-se ver as três portas que a compõem: de esquecimento, de escrita e de saída. A porta de esquecimento recebe a entrada atual e informações de memória e decide que informação será retida e qual será esquecida. A porta de escrita armazena a informação na memória (a depender da saída da porta de esquecimento e da entrada atual da célula) e a porta de saída decide qual será a saída da célula, levando em consideração o estado atual da célula (sua memória) e a entrada atual na mesma [2]. As equações que definem seus valores são

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \tag{2.3}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \tag{2.4}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \tag{2.5}$$

$$\tilde{c} = \sigma_c(W_c x_t + U_c h_{t-1} + b_c), \tag{2.6}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \tag{2.7}$$

e

$$h_t = o_t \circ \sigma_h(c_t), \tag{2.8}$$

onde x_t é o vetor de entrada, f_t é o vetor de ativação da porta de esquecimento, i_t é o vetor de ativação da porta de escrita, o_t é o vetor de ativação da porta de saída, h_t é o vetor de saída da unidade LSTM, $\tilde{c_t}$ é o vetor de ativação da entrada da célula, c_t é o vetor de estado da célula, W e U são matrizes de peso, b é o vetor de bias, σ_g é a função de ativação sigmoidal e σ_c é a função de ativação tangente hiperbólico. x_t tem dimensão d, W tem dimensão $h \times d$, U tem dimensão $h \times h$ e os demais vetores tem dimensão h, com d e h se referindo ao número de características de entrada e número de unidades escondidas, respectivamente [18]. A operação $\circ$ denota o produto de Hadamard, indicando a multiplicação feita de elemento a elemento [19].

2.1.3 Treinamento

Em geral, cada entrada aplicada na rede gera uma saída, e seu valor depende dos valores dos pesos e bias associados a cada neurônio ou célula, assim como da entrada. O processo de treinamento consiste em adaptar esses pesos e bias de forma que a saída fique o mais próximo possível do esperado para uma grande quantidade de dados de entrada.

Para o caso de treinamento supervisionado, durante a etapa de treinamento, cada entrada aplicada é previamente associada a uma saída esperada, denominada *target* ou alvo. O treinamento supervisionado consiste em aplicar cada uma das entradas na rede, comparar a saída obtida com o *target* e adaptar seus pesos e bias de acordo com a diferença, ou erro, entre ambas. Após aplicar todas as entradas de treino, ou, em alguns casos, um conjunto dos dados de treino, diz-se que a rede foi treinada por uma época. Esse processo é repetido diversas vezes, podendo-se simplesmente determinar um número fixo de épocas a treinar, ou configurando algumas condições para que o treino se encerre antes de alcançar esse número fixo. Este é o caso, por exemplo, de a rede já ter alcançado uma boa acurácia mesmo sem ter atingido o número de épocas estabelecido. Após finalizado o treinamento, espera-se que a rede esteja próxima de um mínimo local da função de erro.

2.1.4 Validação cruzada usando K-Fold

Em geral, o objetivo de uma rede neural é fornecer um modelo que represente um sistema de forma generalizada. Desta forma, treinando uma rede com um certo conjunto de dados, a rede deve ser capaz de fornecer uma saída acurada mesmo para dados com que ela nunca teve contato antes. Caso a rede seja capaz apenas de fornecer a saída correta para os dados de treinamento e apresentando desempenho ruim para dados novos, diz-se que a rede está especializada, situação que deseja-se evitar.

Nesse contexto, os métodos de validação cruzada entram para analisar quão generalizada está uma rede treinada, além de permitir uma melhor comparação entre duas redes treinadas com o mesmo conjunto de dados [20].

Ao lidar com as amostras disponíveis para treino de uma rede, uma prática comum é particionar esse conjunto de amostras em dois subconjuntos, um de treinamento e outro de teste. O subconjunto de treino é aplicado na rede enquanto esta é treinada, e seus pesos são ajustados de

acordo com as saídas associadas a esse subconjunto. Quando a rede finaliza seu treinamento, o subconjunto de teste é utilizado para analisar a acurácia da mesma, comparando a saída esperada com a saída obtida. Dessa forma, ao se utilizar um conjunto de amostras isoladas do treinamento, pode-se conferir se a rede aprendeu o modelo geral do sistema, ou se consegue fornecer as respostas corretas apenas para o conjunto com que foi treinada (especialização da rede). Algumas abordagens utilizam ainda um terceiro subconjunto, chamado de validação e utilizado durante o treinamento apenas para conferir e evitar a especialização da rede.

A depender de quais amostras foram separadas para treino e teste, pode ser que haja uma variação da acurácia obtida da rede treinada. Um certo conjunto de teste pode fornecer uma acurácia maior que outro conjunto na mesma rede. Com isto em mente, para garantir maior confiabilidade na acurácia obtida, aplica-se o método de K-Fold.

Pelo K-Fold, o conjunto de amostras é dividido em K subconjuntos de amostras, denominados grupos, mantendo um desses grupos para teste da rede, e os outros K-1 grupos para treino. São treinadas K redes, sendo que em cada vez um grupo diferente é selecionado para teste, e os demais permanecem para treino. Ao final dessas K iterações, chamadas de "rodadas de treinamento", a acurácia da rede será dada pela média das acurácias obtidas em cada um dos treinos realizados. Nota-se que uma prática comumente utilizada é estratificar o conjunto de amostras antes de separá-lo em grupos, isto é, cada grupo deve representar bem o conjunto total [20]. Em um caso de classificação entre duas classes, a proporção dessas amostras em um grupo deve ser próximo à proporção presente no conjunto original.

2.2 Acústica

Tendo em vista que o sinal a ser utilizado como entrada das redes neurais artificiais é uma representação do som, entender suas propriedades acústicas trará benefícios para melhor modelar a rede e avaliar os resultados. Adiante, é apresentada uma forma de representação do espectro de frequências de um sinal sonoro, caracterização da evolução do som ao longo do tempo e descritores sonoros.

2.2.1 Espectro de frequências

Ao analisar sinais que variam ao longo do tempo, uma informação importante que pode ser aferida do sinal são as frequências que compõem o mesmo. Quando o sinal é composto apenas de uma onda senoidal, descobrir a frequência e amplitude da mesma é uma tarefa trivial. Entretanto, para sons mais complexos como o de instrumentos musicais, faz-se necessário o uso de técnicas para recuperar as frequências, e respectivas amplitudes e fases, que compõem o sinal. Uma dessas técnicas é a Transformada de Fourier de Tempo Curto (em inglês, *Short-Time Fourier Transform*), utilizada para a construção de espectrogramas de sinais digitais [21]

2.2.1.1 Short-Time Fourier Transform (STFT)

A STFT é relacionada com a Transformada de Fourier (FT, do inglês *Fourier Transform*). Ao se aplicar a FT em um sinal variante no tempo $x(t)$, perde-se a informação temporal, mas obtemos um sinal $X(\omega)$ em função da frequência, informando quais frequências, suas amplitudes e fases estão presentes no sinal.

Para o caso da STFT, o sinal é particionado em trechos menores (janelas), e uma variação da FT para tempo discreto é aplicada em cada um desses trechos [21] [22]. Concatenando trechos subsequentes, é possível analisar a evolução ao longo do tempo do espectro de frequências do sinal, isto é, das frequências presentes em cada trecho.

2.2.1.2 Espectrograma

Um espectrograma pode ser considerado uma representação visual do espectro de frequências adquirido pela STFT com os parâmetros frequência, tempo e amplitude, e rejeitando informações de fase [23]. A Figura 2.6 mostra o exemplo de uma onda senoidal de 440 Hz, com sua forma de onda em cima e o respectivo espectrograma embaixo, obtidos com o Sonic Visualizer [24], software desenvolvido por Cannam *et al.* [25] em 2010.

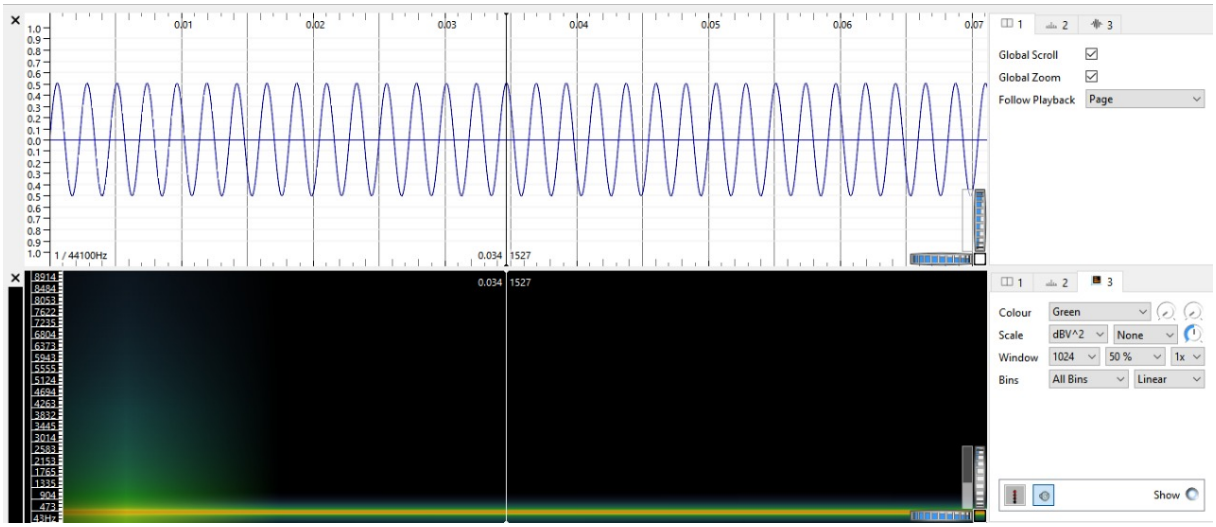
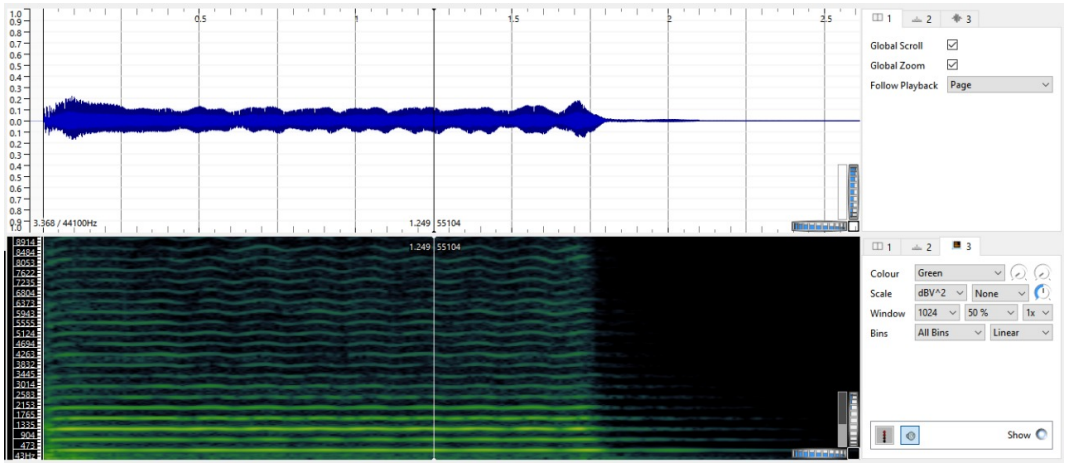


Figura 2.6: Representação da forma de onda e espectrograma de um sinal senoidal de 440 Hz.

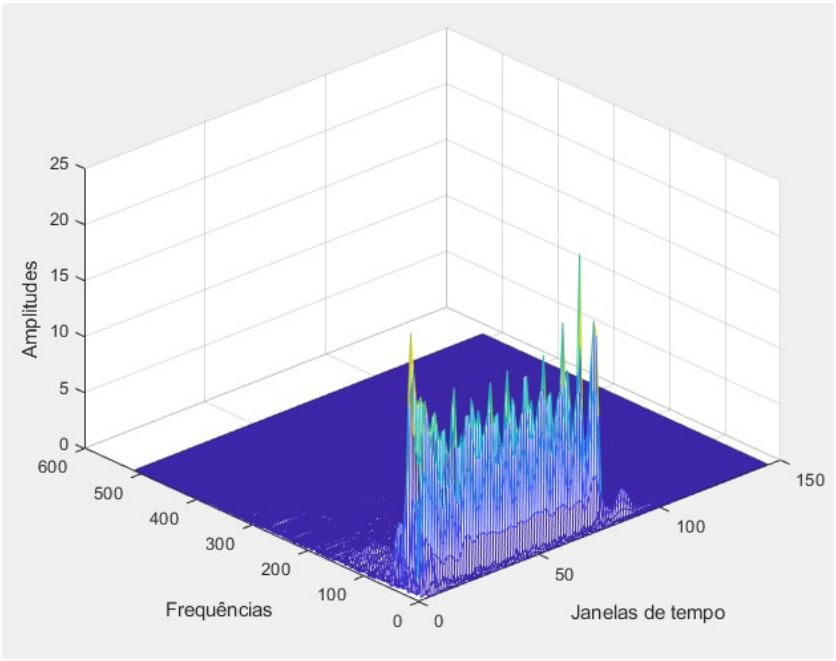
Como apresentado pelo gráfico, em um espectrograma o eixo das abscissas traz a informação temporal e é composto pelos trechos em que o sinal foi reparticionado para realizar a STFT. O eixo das ordenadas indica as frequências, a amplitude de cada frequência é indicada por um padrão de cores. No exemplo apresentado, o padrão de cores é o de cores frias e quentes, indo do preto para amplitudes nulas até o vermelho para altas amplitudes. Nota-se também o painel na direita do espectrograma, utilizado para ajustar alguns parâmetros de visualização. Destaca-se aqui o uso da escala dBV^2, fazendo com que o mapeamento das cores seja proporcional ao logaritmo do quadrado da amplitude absoluta [26].

Analisando o espectrograma, pode-se ver que em toda sua extensão, a faixa de frequências ao redor de 440Hz apresenta uma cor mais quente, como era de se esperar para uma senoide desta frequência. Além disso, como a amplitude da senoide se mantém constante ao longo do tempo, observa-se que a cor também se mantém constante. Apenas no começo se observa a representação de frequências fora de 440 Hz, mas esta anomalia está presente apenas na primeira e na última janela temporal do espectrograma.

Para exemplificar o espectrograma de um instrumento musical, é apresentada a Figura 2.7a. O sinal é de um contra-baixo tocando a nota Lá 4, com frequência fundamental em 440Hz, sendo o gráfico superior a onda ao longo do tempo, e o inferior, o espectrograma do mesmo. Note que a escala de tempo é maior que a apresentada na Figura 2.6. Para fins de comparação, a Figura 2.7b apresenta o espectrograma do mesmo sinal, mas em uma visualização 3D.



(a) Forma de onda e espectrograma no padrão de cores frias/quentes.



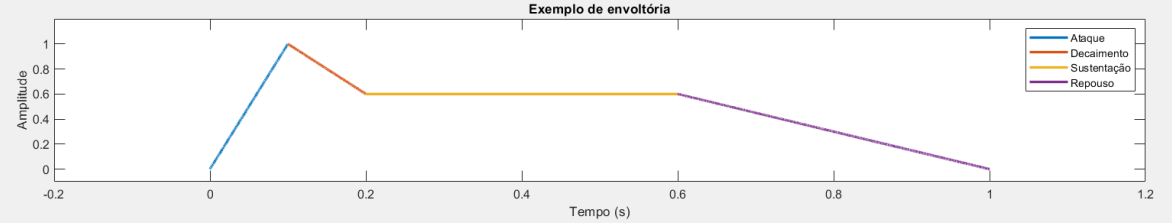
(b) Espectrograma 3D.

Figura 2.7: Forma de onda e espectrogramas da nota Lá 4 sendo tocada em um contrabaixo.

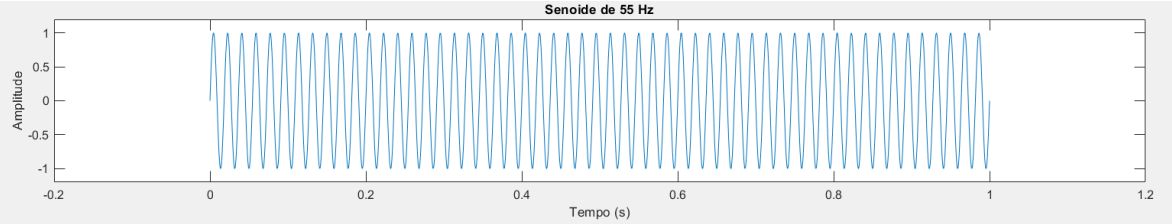
Nestes espectrogramas percebe-se mais a variação da amplitude ao longo do tempo, e menos da onda em si. O espectrograma fornece algumas informações importantes a respeito de sinais de áudio de instrumento. Uma digna de nota é que, apesar de haver componentes de muitas outras frequências, as mais marcantes correspondem à fundamental (440Hz) e seus harmônicos, isto é, as frequências múltiplas da fundamental 880Hz, 1320Hz e assim por diante.

2.2.2 Envoltória

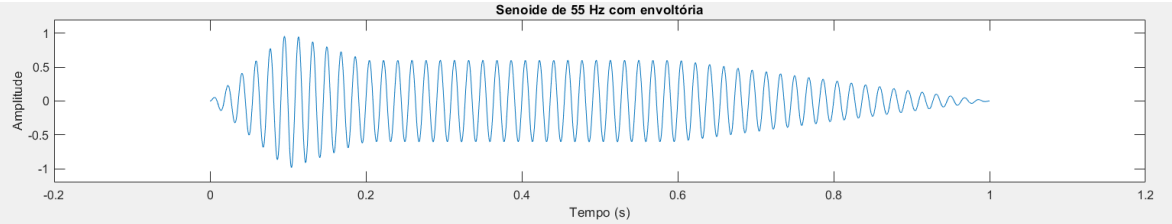
A envoltória de um som está relacionada com o seu desenvolvimento, principalmente da amplitude da onda, ao longo do tempo em que a nota é tocada. De forma genérica, a envoltória de instrumentos pode ser dividida em Ataque, Decaimento, Sustentação e Repouso (ADSR, também do inglês *Attack*, *Decay*, *Sustain* e *Release*), apesar de que, a depender do instrumento, alguma dessas fases pode não estar presente. O ataque consiste no momento inicial em que o instrumento é tocado, nas alterações que ocorrem até que o som alcance a intensidade de estado estacionário. A sustentação se refere ao período em que a intensidade do som se mantém relativamente constante. O decaimento é a queda de intensidade que pode ocorrer entre os momentos de ataque e sustentação. O repouso é o período após a sustentação em que a intensidade do som decresce continuamente até o silêncio. Em alguns contextos, são apenas utilizados os termos ataque, sustentação e decaimento, sendo neste caso o decaimento equivalente ao que fora definido como repouso. A Figura 2.8 mostra um exemplo de onda com envoltória.



(a) Envoltória.



(b) Onda senoidal.



(c) Onda senoidal com envoltória.

Figura 2.8: Exemplo de envoltória aplicada em onda senoidal.

A sub figura 2.8a mostra o formato da envoltória aplicada. No gráfico, os trechos correspondentes ao ataque, decaimento, sustentação e repouso são representados pelas retas em azul, vermelho, amarelo e roxo, respectivamente. Multiplicando esta envoltória pela senoide de 55Hz da sub figura 2.8b, obtemos a onda com a envoltória apresentada, cujo resultado pode ser visualizado na sub figura 2.8c.

É importante notar que a envoltória apresentada foi criada artificialmente, e que sinais de áudio criados por instrumentos apresentam uma envoltória própria naturalmente. Tomando como exemplo o sinal de contra-baixo apresentado na Figura 2.7a, pode-se especular sua envoltória como apresentado na Figura 2.9.

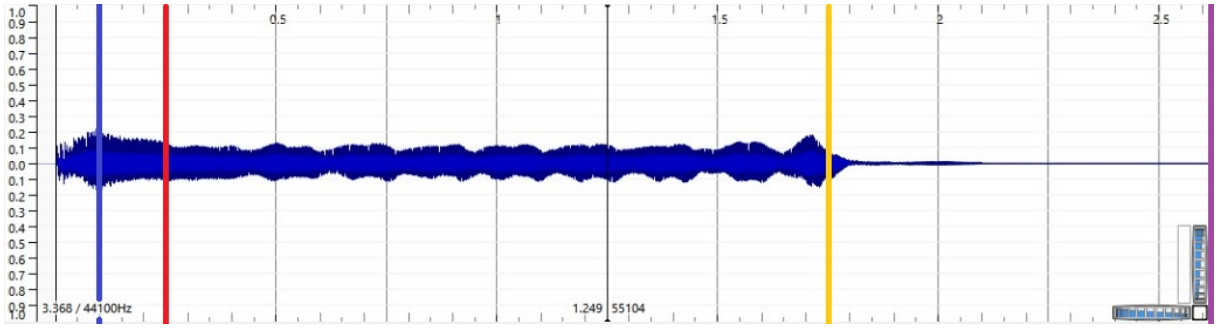


Figura 2.9: Períodos de ADSR da onda apresentada na Figura 2.7a.

Na Figura apresentada, pode-se especular que o ataque do instrumento esteja no período de 0 a 0,1s, o decaimento entre 0,1 e 0,25s, com uma sustentação indo desde 0,25 até 1,75s e o repouso desde este instante até o final do áudio. Além da variação da intensidade do volume, existem outras informações importantes do sinal que ocorrem em trechos específicos da envoltória e que auxiliam na identificação do instrumento. Pelo espectrograma da Figura 2.7a, nota-se que durante o ataque da nota, há uma presença mais forte de frequências que não são harmônicos da frequência principal, e no repouso, muitos harmônicos superiores perdem intensidade rapidamente.

2.2.3 Descritores sonoros

Descritores sonoros, como o nome indica, são parâmetros para descrever um sinal sonoro a partir de características extraídas do mesmo, a maioria de caráter estatístico. Alguns descritores que podem ser extraídos do espectro de frequências de um som são o centroide, o espalhamento, o declive, o decrescimento e o ponto de *roll-off* espectral.

2.2.3.1 Centroide Espectral

O Centroide Espectral é o baricentro do espectro [27]. Ele é calculado como

$$\mu = \frac{\sum_{i=0}^{n-1} f[i]a[i]}{\sum_{i=0}^{n-1} a[i]} \tag{2.9}$$

onde μ é o Centroide Espectral, em Hertz, n é metade do tamanho da janela, i é o índice do vetor que compõe o sinal no domínio da frequência, $a[i]$ é a amplitude do sinal na frequência indicada pelo índice i, na parte real do cálculo da FFT (do inglês, *Fast Fourier Transform*), e

$$f[i] = i \cdot \frac{\text{frequência de amostragem}}{\text{tamanho da janela da FFT}} \tag{2.10}$$

Sendo o baricentro do espectro, pode-se afirmar que além de ser influenciado pela frequência fundamental da nota, o Centroide Espectral também sofre alterações em seu valor a depender da intensidade de seus harmônicos superiores e do ruído presente no sinal.

2.2.3.2 Espalhamento Espectral

O Espalhamento Espectral é definido como a variância do Centroide Espectral [27], e pode ser calculado como

$$\nu = \frac{\sum_{i=0}^{n-1}(f[i] - \mu)^2 a[i]}{\sum_{i=0}^{n-1} a[i]} \tag{2.11}$$

onde ν é o Espalhamento Espectral. Assim como o Centroide Espectral, ν é influenciada pela intensidade dos harmônicos superiores e pelo ruído. Fornece informações de quão esparsas estão distribuídas as frequências do sinal.

2.2.3.3 Declive Espectral

Declive Espectral é uma estimação da quantidade de magnitude espectral decrescendo [27], calculada por

$$slope = \frac{1}{\sum_{i=0}^{n-1} a[i]} \frac{n\sum_{i=0}^{n-1} f[i]a[i] - \sum_{i=0}^{n-1} f[i] \sum_{i=0}^{n-1} a[i]}{n\sum_{i=0}^{n-1} f^2[i] - (\sum_{i=0}^{n-1} f[i])^2}. \tag{2.12}$$

2.2.3.4 Decrescimento Espectral

É semelhante ao Declive Espectral, representando a quantidade de magnitude espectral decrescendo, mas em tese é mais correlacionado à percepção humana [27]. É definido por

$$decrease = \frac{\sum_{i=0}^{n-1} a[i] - a[1]}{\sum_{i=0}^{n-1} a[i](i-1)}. \tag{2.13}$$

2.2.3.5 Ponto de Roll-Off Espectral

O ponto de *Roll-Off* Espectral é a frequência $f_c[i]$ em que uma porcentagem do sinal fica abaixo desta frequência, por padrão sendo utilizado 95% do sinal [27]. O ponto de *Roll-Off* é calculado por

$$\sum_{i=0}^{f_c[i]} a^2[f[i]] = x \sum_{i=0}^{n-1} a^2[f[i]] \tag{2.14}$$

onde $f_c[i]$ é o ponto de *Roll-Off* e x é a porcentagem de energia *Roll-Off* acumulada.

2.2.3.6 Vetor de Descritores

Neste trabalho, um Vetor de Descritores X_d é definido a partir dessas medidas a fim de caracterizar um trecho de n amostras (janela) de um sinal de áudio, na forma

$$X_d(k) = \begin{bmatrix} \mu \\ \nu \\ slope \\ decrease \\ f_c[i] \end{bmatrix} \tag{2.15}$$

O vetor coluna X_d possui dimensão 5 e caracteriza o k-ésimo trecho de sinal de áudio.

2.3 Métricas de avaliação de desempenho

Ao desenvolver e propor um novo sistema para realizar uma determinada tarefa, como a de classificar e a de separar fontes sonoras, é importante que haja alguma forma padronizada de avaliar o desempenho do mesmo. Deste modo, é possível comparar sistemas diferentes, cujo objetivo seja o mesmo, e analisar numericamente qual apresenta melhores resultados.

Para problemas de classificação, em que cada entrada é associada a uma classe específica, uma forma simples de avaliar a acurácia do sistema é através da porcentagem de acertos em relação ao conjunto de teste total. Por exemplo, em um sistema em que as entradas devem ser associadas às duas classes “a” e “b”, caso o conjunto de teste consista em 100 entradas e o sistema classifique corretamente 90 destas, sua acurácia será de 90%.

Entretanto, esta métrica é muitas vezes considerada simples para avaliar o desempenho [28]. Informações como “quantas vezes uma entrada da classe ’a’ foi associada à classe ’b’ ” acabam se perdendo. Com isso em mente, outra métrica que pode ser utilizada é a partir de uma matriz de confusão. Um exemplo de matriz de confusão pode ser visto na Figura 2.10. Nela, encontra-se a informação de quantas vezes uma classe prevista pela rede é associada a cada uma das classes esperadas. A informação de que 10 vezes o sistema confundiu que uma entrada da classe “b” fosse da classe “a”, e nenhuma vez confundiu uma entrada da classe “a” com classe “b”, é importante em alguns casos para uma análise mais profunda do mesmo, e esta é explícita em uma matriz de confusão.

		Classe esperada	
		a	b
Classe prevista	a	50	10
	b	0	40

Figura 2.10: Exemplo de matriz de confusão com classes “a” e “b”.

Para o caso da separação de fonte sonora, Vincent *et al.* [29] propõem que a avaliação de uma separação de fonte de áudio seja feita com base em quatro parâmetros: *Source to Distortion Ratio* (SDR), *Source to Interferences Ratio* (SIR), *Sources to Noise Ratio* (SNR) e *Sources to Artifacts Ratio* (SAR). Para encontrar tais valores, o sinal estimado é primeiramente decomposto conforme

$$\hat{s} = s_{target} + e_{interf} + e_{noise} + e_{artif}, \tag{2.16}$$

onde $\hat{s}$ é o sinal estimado, s_{target} é o sinal esperado modificado por alguma distorção permitida, e_{interf} corresponde a interferências vindas de outras fontes sonoras, e_{noise} é a componente de ruído presente no sinal obtido e e_{artif} são os artefatos presentes no sinal de saída.

A partir dessas componentes, os valores de SDR, SIR, SNR e SAR são definidos por

$$SDR := 10log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf} + e_{noise} + e_{artif}\right\|^2}, \tag{2.17}$$

$$SIR := 10log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf}\right\|^2}, \tag{2.18}$$

$$SNR := 10log_{10}\frac{\left\|s_{target} + e_{interf}\right\|^2}{\left\|e_{noise}\right\|^2}, \text{ e} \tag{2.19}$$

$$SAR := 10log_{10}\frac{\left\|s_{target} + e_{interf} + e_{noise}\right\|^2}{\left\|e_{artif}\right\|^2}. \tag{2.20}$$

Para o sinal completo, este é dividido em janelas, os parâmetros são medidos para cada uma delas, e o valor resultante do sinal é dado pela média obtida em cada janela.

Pode-se interpretar a SAR como uma medida da quantidade de artefatos que a fonte estimada possui, SNR como a quantidade de ruído, SIR indica o quanto de outras fontes estão presentes no sinal separado, o que se aproxima do conceito de *leakage* ou vazamento, e SDR pode ser

considerado como uma medida geral de quão bom está a estimação da fonte. Em artigos que apresentam apenas um destes parâmetros, comumente é utilizado o valor SDR [30].

Nota-se que esta forma de avaliação é objetiva e necessita apenas que se tenha previamente a fonte separada verdadeira e o nível de distorção aceitável [29]. Existem métodos que se utilizam de avaliação humana e que fornecem uma métrica mais aceitável por conta disto, porém estes são muito custosos (recursos humanos e financeiros) [30], então não serão abordados neste trabalho.

2.4 Trabalhos relacionados

Através de comparações com outros sistemas propostos para o mesmo objetivo, é possível analisar o quão bom um novo sistema é, utilizando as mesmas métricas de avaliação em um mesmo conjunto de amostras. A seguir, serão apresentados alguns trabalhos relacionados aos temas de classificação e de separação de fonte sonora, respectivamente.

2.4.1 Classificação de Fonte Sonora

Em 2017, Bahre *et al.* [7] propuseram o treinamento de uma Máquina de Vetores de Suporte (do inlgês *Support Vector Machine*) (SVM) de múltiplas classes a partir de três características extraídas das amostras sonoras: inclinação do ataque da envoltória do som, *Constant-Q Transform* (CQT) e coeficientes cepstrais. A classificação fora feita em amostras de flauta, violão, piano, trompete e violino, de um banco de dados de instrumentos musicais do *University of Iowa Electronic Music Studio*, e obteve uma acurácia de 86,81%.

Em 2018, Jeyalakshmi *et al.* [3] fizeram experimentos tanto com o classificador *K-Nearest Neighboor* (K-NN) quanto com *Hidden Markov Models* (HMM), comparando o resultado de ambos a depender de qual característica do sinal era extraída, podendo ser *Mel Frequency Cepstral Coefficients* (MFCC), *Perceptual Linear Prediction* (PLP) ou RASTA-PLP. A classificação fora feita em amostras de flauta, violão, violino e piano de três bancos de dados, realizando treinamentos tanto com apenas a mesma nota sendo tocada quanto com notas diferentes. As acurácias obtidas por estes sistemas podem ser visualizados nas Tabelas 2.1 e 2.2.

Tabela 2.1: Acurácias obtidas em [3] com classificador HMM.

Extração de características	Acurácia (%)	
	Notas iguais	Notas diferentes
MFCC	96.75	90.50
PLP	92	86.75
RASTA-PLP	87	82

Tabela 2.2: Acurácias obtidas em [3] com classificador K-NN.

Extração de características	Acurácia (%)	
	Notas iguais	Notas diferentes
MFCC	85.5	75
PLP	82	74.5
RASTA-PLP	70.50	63

Pelas tabelas, algo que fica claro é que, independentemente do classificador utilizado, o treino com notas iguais obtém sempre um desempenho melhor que o treino com notas diferentes, provavelmente por conta de algum grau de especialização na frequência da nota. Nota-se também que o classificador HMM com extração de características MFCC possui melhor acurácia para os casos de notas iguais e diferentes.

Em 2017, Anderson *et al.* [8] desenvolveram uma rede MLP para classificar os sinais de entrada entre trompete ou tuba a partir de 19 características extraídas dos mesmos. Foram utilizadas amostras do *University of Iowa Musical Instrument Samples database*, sendo utilizados 12 amostras para teste do sistema (6 de cada instrumento), obtendo uma acurácia global de 75%.

2.4.2 Separação de Fonte Sonora

Em 2018, Stoller *et al.* [9] propuseram a Wave-U-Net, uma adaptação da U-Net voltada para sinais de áudio no formato de onda. Trata-se uma rede neural capaz de realizar a separação tanto de sinais monaurais (som captado por apenas um microfone) quanto de sinais estéreo (dois microfones) e fora aplicada tanto para a separação de múltiplos instrumentos do banco de dados MUSDB18 [31] quanto para a separação vocal do banco de dados CCMixter [32].A partir da estrutura básica proposta, foram feitas alterações para analisar que melhorias eram trazidas aos resultados.

Stoller *et al.* [9] também apresentam um problema com a avaliação pelo SDR presente principalmente em separação de vocal. Trechos em que o sinal é silencioso ou muito próximo disso, o resultado de SDR é indefinido ($log(0)$) ou muito baixo, mesmo que a separação tenha sido boa, e o SDR de tais trechos afetam a média geral da separação. Por conta disto, seus resultados apresentam também outras medidas estatísticas que atenuam esta situação, sendo elas a mediana, a mediana do desvio absoluto (do inglês *Median Absolute Deviation*) (MAD) e o desvio padrão (do inglês *Standard Deviation*) (SD) do valor do SDR das janelas do sinal separado. Os resultados obtidos estão nas Tabelas 2.3 e 2.4 , em que M1 corresponde a estrutura básica da Wave-U-Net, M2 aplica a regra de que a somatória dos sinais separados deve ser igual ao sinal mixado, M3 insere entrada de contexto, M4 realiza a separação de sinais estéreo (os demais são monaurais), M5 aplica uma nova camada de *upsampling*, M6 corresponde à aplicação da M4 em separação multi-instrumental (as demais fazem a separação vocal). U7 e U7a correspondem ao treinamento de uma rede U-Net baseada em espectrograma e M7 é uma rede Wave-U-Net baseada na M4, todas nas mesmas condições para que seja possível compará-las.

Tabela 2.3: Resultado da separação vocal das redes Wave-U-Net e U-Net.

		M1	M2	M3	M4	M5	M7	U7	U7a
Vocal	Mediana	3.90	3.92	3.96	4.46	4.58	3.49	2.76	2.74
	MAD	3.04	3.01	3.00	3.21	3.28	2.71	2.46	2.54
	Média	-0.12	0.05	0.31	0.65	0.55	-0.23	-0.66	0.51
	SD	14.00	13.63	13.25	13.67	13.84	13.00	12.38	10.82
Acomp.	Mediana	7.45	7.46	7.53	10.69	10.66	7.12	6.76	6.68
	MAD	2.08	2.10	2.11	3.15	3.10	2.04	2.00	2.04
	Média	7.62	7.68	7.66	11.85	11.74	7.15	6.90	6.85
	SD	3.93	3.84	3.90	7.03	7.05	4.10	3.67	3.60

Tabela 2.4: Resultado da separação multi-instrumental da rede Wave-U-Net.

Fonte	Mediana	MAD	Média	SD
Vocal	3.0	2.76	-2.10	15.41
Baixo	2.91	2.47	-0.30	13.50
Bateria	4.15	1.99	2.88	7.68
Outros	2.03	1.64	1.68	6.14

Alguns pontos a se destacar das tabelas: na comparação entre Wave-U-Net e U-Net, a primeira apresenta resultados melhores; a separação de sinais estéreo e a aplicação da nova camada de *upsampling* apresentam os melhores resultados; na separação multi-instrumental, percebe-se o problema informado do SDR em sinais vocais, com a pior média dentre as fontes, mas a segunda maior mediana.

Em 2019, Dèfossez *et al.* [4] propuseram o separador Demucs, uma rede neural de arquitetura encoder-decoder, com camadas de encoder convolucionais, LSTM bidirecionais e decoder convolucional. Demucs é utilizado para separação de sinais estéreo, utilizando como entrada a forma de onda do sinal e fornecendo de saída as fontes estimadas na forma de onda. O sistema foi testado no banco de dados MusDB [31] para separação multi-instrumental, e os resultados podem ser visualizados na Tabela 2.5, comparando com outros sistemas de separação de fonte sonora.

Tabela 2.5: Comparação do Demucs com outros separadores de fonte sonora [4].

			SDR em dB				
Arquitetura	**Onda?**	**Extra?**	**Todos**	**Bateria**	**Baixo**	**Outros**	**Vocais**
IRM oracle	Não	N/A	8.22	8.45	7.12	7.85	9.43
Open-Unmix	Não	Não	5.33	5.73	5.23	4.02	6.32
Wave-U-Net	Sim	Não	3.23	4.22	3.21	2.25	3.25
Demucs	Sim	Não	5.58	6.08	5.83	4.12	6.29
Conv-Tasnet	Sim	Não	5.73	6.08	5.66	4.37	6.81
Demucs	Sim	150	6.33	7.08	6.70	4.47	7.05
Conv-Tasnet	Sim	150	6.32	7.11	7.00	4.44	6.74
MMDenseLSTM	Não	804	6.04	6.81	5.40	4.80	7.16

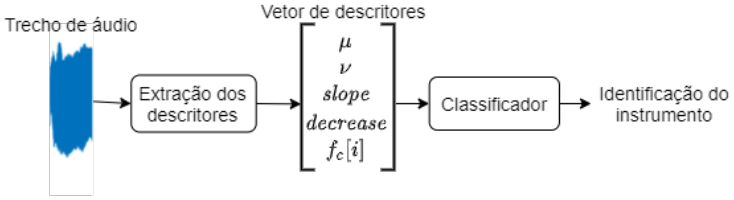
A coluna “Onda?” indica se a sistema atua no sinal na forma de onda (sim) ou no domínio da frequência (não) e “Extra?” indica se e quantas músicas extras foram utilizadas para treinamento da rede. A comparação é feita com a mediana do valor de SDR, e a coluna “Todos” apresenta a média obtida para os quatro instrumentos. Acrescenta-se que a arquitetura ”Conv-Tasnet” apresentada na tabela é uma adaptação desta rede feita por Dèfossez *et al.* [4] para aceitar sinais estéreos, fator que incrementa sua avaliação no SDR. Nota-se que esta possui um resultado melhor que a Demucs quando treinados sem músicas extras, mas ao aplicar a mesma quantidade de músicas extras, Demucs fica comparável a ela.

Capítulo 3

Metodologia

Este capítulo é destinado a explicitar como os sistemas de classificação e separação foram feitos, apresentando os programas e o banco de amostras utilizados, pré-processamentos realizados nos sinais e arquitetura e treinamento das redes desenvolvidas. Os sistemas possuem pequenas diferenças nas etapas de treinamento da rede e na aplicação das mesmas. A Figura 3.1 apresenta os esquemáticos do sistema de classificação em cada etapa.

(a) Etapa de treinamento.



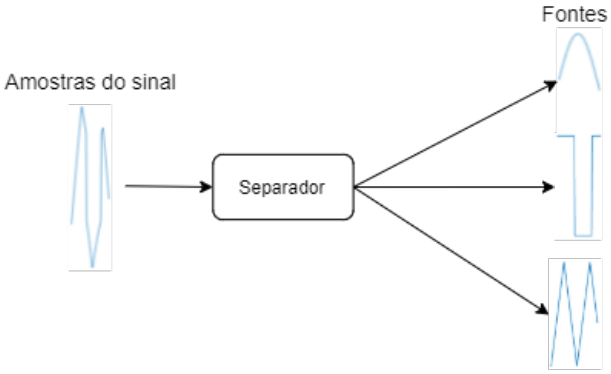
(b) Etapa de aplicação.

Figura 3.1: Esquemáticos do sistema de classificação de fonte sonora.

Na etapa de treinamento (Figura 3.1a), o sistema recebe um sinal de áudio completo. Este é pré-processado com a retirada de seus trechos de silêncio e com a obtenção do vetor de descritor de cada trecho de áudio. O classificador é treinado com cada um dos vetores obtidos (ou sequência de vetores) para classificar o instrumento a cada trecho do áudio. Na etapa de aplicação (Figura 3.1b), o sinal é obtido continuamente, sendo extraído um vetor ou pequena sequência de descritores a cada trecho do sinal, e a classificação é feita nesse trecho do áudio recém-obtido.

Os sistemas desenvolvidos para separação de fonte sonora são representados pelos esquemáticos da Figura 3.2.

(a) Etapa de treinamento.



(b) Etapa de aplicação.

Figura 3.2: Esquemáticos do sistema de separação de fonte sonora.

De forma semelhante ao sistema de classificação, um sinal completo de áudio é fornecido na etapa de treinamento, com os momentos de silêncio sendo retirados no pré-processamento. O sinal resultante é fornecido de amostra a amostra para treinamento da rede, sendo importante que a sequência temporal das amostras seja mantida. Para cada amostra, são fornecidas as componentes de cada fonte que compõe a entrada em canais diferentes. Na etapa de aplicação, para cada amostra obtida de algum sensor são fornecidas as componentes de cada fonte que compõe a entrada, através da rede treinada.

Nas seções seguintes serão detalhadas cada parte dos sistemas, assim como do ambiente de desenvolvimento.

3.1 Programas utilizados

Para a aquisição e condicionamento dos sinais de áudio utilizados, desenvolvimento das redes neurais propostas e análise dos resultados, foram desenvolvidos códigos utilizando o software MATLAB, da MathWorks [33], na versão 2021a. Dentre as toolboxes utilizadas, destacam-se a Audio Toolbox e a Deep Learning Toolbox. Os códigos desenvolvidos podem ser encontrados no repositório deste projeto no GitHub [34]

Para análise do espectro de frequências dos sinais através de espectrogramas, foi utilizado o software Sonic Visualizer [24][25]. Dentre as configurações apresentadas no painel do espectro-

grama, destaca-se que a escala fora colocada para dbV^2 e o limiar (*threshold*) para $-\infty$. Todos os espectrogramas do Capítulo 4 - Resultados - apresentam essa mesma configuração, alterando apenas a escala do tempo e a escala de frequência, para destaque de frequências inferiores, quando não há informação útil nas altas frequências.

3.2 Banco de sinais instrumentais

O objetivo deste trabalho é classificar e separar sinais produzidos por instrumentos musicais. Para isto, foram utilizados sinais de áudio da coleção "IRCAM solo instruments 2" [35]. Cada sinal possui o som produzido por um instrumento individualmente. Além do instrumento, existem ainda algumas outras informações sobre o áudio, tais como

- *Técnica de articulação:* Forma com que o instrumento é tocado, o que afeta o som produzido.
- *Intensidade:* O instrumento do sinal pode ser tocado com intensidades diferentes, variando entre pianíssimo (pp, intensidade fraca), mezzo forte (mf, intensidade intermediária) e fortíssimo (ff, intensidade forte).
- *Corda:* Exclusivo para instrumentos de cordas, uma mesma altura pode ser tocada em cordas diferentes, fator que afeta o som produzido.
- *Nota:* Nota musical tocada pelo instrumento.

Esse banco de dados possui áudios de 16 instrumentos orquestrais diferentes. Para cada instrumento, são inclusas as formas padrão de se tocar e inúmeras articulações avançadas e experimentais [35], podendo tocar com as diferentes intensidades e explorando toda a extensão de notas dos instrumentos. A Figura 3.3 mostra o exemplo de um áudio desse banco.

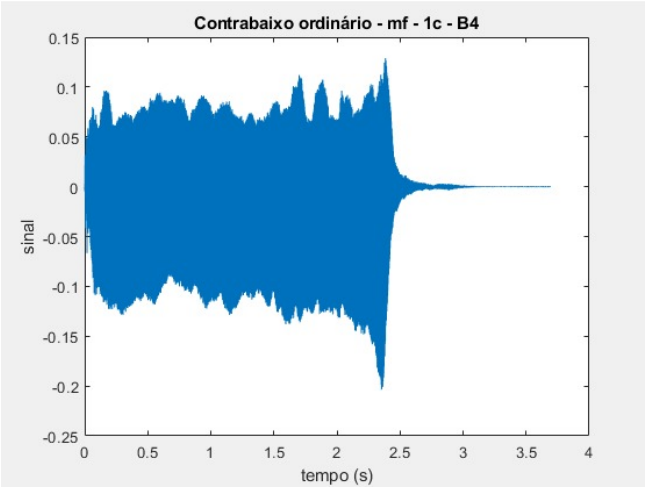


Figura 3.3: Exemplo de sinal instrumental obtido do "IRCAM solo instruments 2".

O sinal é de um contrabaixo tocado de forma ordinária e mezzo forte, na primeira corda e com a nota B4.

3.3 Extração do silêncio

Observando o exemplo de áudio da Figura 3.3, percebe-se um problema de utilizar o sinal em sua forma pura. Em sua duração total, há períodos em que o instrumento não é tocado. Estes períodos não podem ser inseridos como treinamento da rede, para que a mesma não identifique silêncio como sinal de um instrumento. Assim, faz-se necessário realizar um pré-processamento em cada sinal de áudio para retirar o silêncio.

Uma das formas utilizadas para separar o silêncio do som instrumental utiliza uma adaptação da segmentação de silêncio no domínio do tempo proposta por Zhang *et al.* [36]. Na segmentação proposta, o sinal é dividido em janelas de tamanho igual e encontra-se a amplitude absoluta média do sinal em cada uma das janelas. A partir dessas amplitudes, calcula-se um limiar por meio de

$$L = \begin{cases} 0{,}015 \cdot \max(A), & \text{se média}(A) > 0{,}075 \cdot \max(A) \\ 0{,}005 \cdot \max(A), & \text{se média}(A) \leq 0{,}075 \cdot \max(A) \end{cases} \tag{3.1}$$

onde L é o valor do limiar e A se refere ao conjunto das amplitudes absolutas médias de cada janela.

Uma janela será considerada o início de um segmento com som caso esta tenha amplitude absoluta média acima do limiar e as duas janelas anteriores a si tenham este valor abaixo do limiar. Uma janela será considerada o final de um segmento com som caso esta tenha amplitude absoluta média acima do limiar e as duas janelas posteriores a si tenham esse valor abaixo do limiar. Dessa forma, pode-se segmentar um sinal de áudio que possua mais de um trecho de silêncio e som. Considerando que deve haver apenas um sinal instrumental no áudio, este método é adaptado mantendo apenas a primeira janela de início e de final encontradas. O sinal pré-processado final consistirá das amostras entre tais janelas.

Além dessa forma automática de segmentar o sinal, outro método aplicado foi pela escolha arbitrária dos pontos de início e fim do sinal, auxiliado pela observação da forma de onda. O sinal resultante consiste apenas das amostras entre os pontos de início e fim.

3.4 Classificação baseada em Descritores

Nesta etapa, serão propostos sistemas de classificação de fonte sonora baseados nos descritores apresentados na Seção 2.2.3. Os sinais são segmentados em janelas de tamanho n amostras, os cinco descritores de cada segmento são calculados e aplicados aos classificadores na forma de uma sequência de Vetores de Descritores.

Um dos modelos avaliados para resolução de tal problema são redes MLP com 1 camada escondida, conforme exemplificado na Figura 3.4.

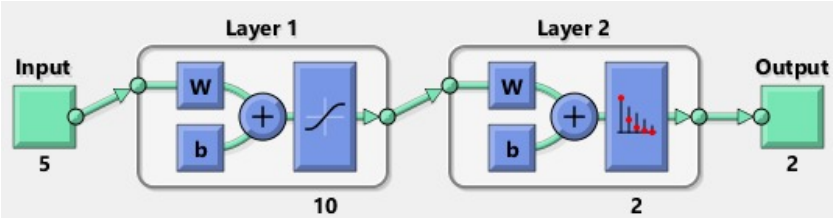


Figura 3.4: Rede MLP usada na classificação por descritores.

A entrada é dada pelo Vetor de Descritores, conferindo à rede a capacidade de classificar o instrumento em trechos do sinal de áudio, e cada uma de suas saídas, com função de ativação *softmax*, representa a probabilidade de a entrada pertencer a um dos instrumentos, variando de 0 a 1. A quantidade de saídas da rede é igual à quantidade de instrumentos que ela foi treinada para classificar. São feitos testes variando a quantidade de neurônios na camada intermediária.

Outra rede avaliada para a classificação é a rede recorrente de Elman, mostrada na Figura 3.5.

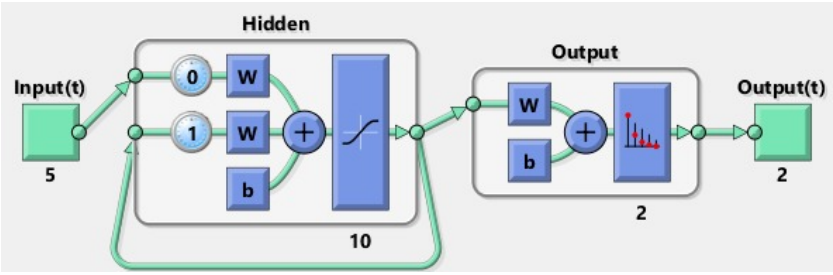


Figura 3.5: Rede Elman usada na classificação por descritores.

Sendo uma rede recorrente, sua entrada deve ser uma sequência temporal. Para tanto, a sequência de Vetores de Descritores é apresentada às entradas da rede. A sequência deve corresponder a um trecho de aproximadamente 0,1 segundos, tempo escolhido arbitrariamente visando permitir a classificação online do sinal. Assim, considerando a frequência de amostragem do sinal em 44,1 kHz e que cada vetor de descritores possui 1024 amostras, cada sequência de vetores é composta por quatro vetores de descritores subsequentes, representando um trecho de 0,0929 segundos do sinal. Assim como feito na MLP, a quantidade de neurônios na camada intermediária da rede de Elman é variada em cada teste, e a saída é composta por uma camada *softmax*.

O objetivo é verificar o desempenho de uma rede que avalia os vetores da sequência de modo individual (MLP) com aquela que avalia o comportamento baseado na sequência temporal dos vetores (Elman).

3.5 Separação na forma de onda

Nesta etapa, serão propostos sistemas que realizem a separação de fonte sonora. Inserindo um sinal composto por múltiplas fontes sonoras, tal sistema deve fornecer como saída cada uma dessas fontes separadamente.

Cada sinal mixado é obtido pela mistura de N_F fontes através de

$$m(t) = \frac{1}{N_F} \cdot \sum_{i=1}^{N_F} f_i(t), \tag{3.2}$$

onde $m(t)$ é o sinal com as fontes misturadas no tempo t, N_F é o número de fontes presentes na mistura e $f_i(t)$ é o sinal da fonte i no tempo t. Após a separação do sinal $m(t)$ nas fontes estimadas $f_{ei}(t)$, espera-se que $f_{ei}(t)$ se aproxime de $\frac{f_i(t)}{N_F}$, para $i = 1...N_F$.

As redes neurais recorrentes de Elman e LSTM serão avaliadas para realização da separação das fontes a partir das amostras do sinal mixado no domínio do tempo. A Figura 3.6 apresenta a rede de Elman desenvolvida.

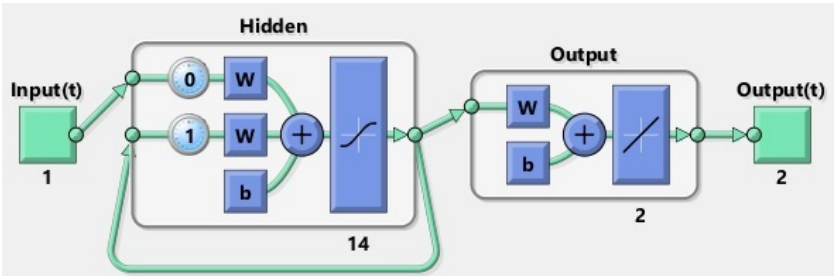


Figura 3.6: Rede Elman usada na separação na forma de onda.

Na rede apresentada, a quantidade de neurônios na camada escondida (*Hidden*) é variada para estudar a capacidade de separação da rede. A quantidade de neurônios na camada de saída (*Output*) é igual à quantidade de fontes usadas nos testes.

A Figura 3.7 apresenta a rede LSTM desenvolvida para separação, onde a quantidade de células de memória na camada escondida é variada para estudar a capacidade de separação da rede, assim como é feito na rede de Elman.

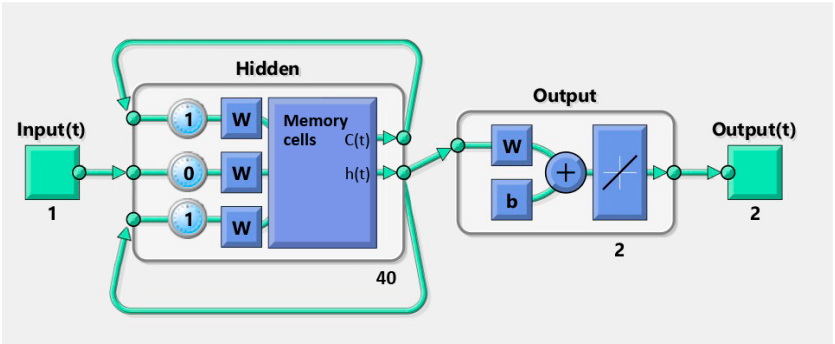


Figura 3.7: Rede LSTM usada na separação na forma de onda.

O número de neurônios na camada de saída também é dependente de quantas fontes são utilizadas para formar a mistura de entrada. Nota-se que para ambas redes a entrada é de apenas uma amostra, variando ao longo do tempo, e os neurônios de saída possuem ativação linear, permitindo saídas no formato de onda.

Capítulo 4

Resultados

Neste capítulo serão apresentadas as condições dos testes realizados, junto com os resultados obtidos e uma análise objetiva dos mesmos. Os conteúdos abordam as características do hardware empregado, os sinais utilizados e os pré-processamentos realizados neles, a classificação de sinais instrumentais em uma ou mais notas e a separação das fontes de uma mistura sonora, tanto para sinais artificiais quanto para sinais instrumentais.

4.1 Hardware

Como alguns dos resultados apresentam informações como tempo decorrido e este dado é muito dependente do hardware utilizado, faz-se necessário apresentá-lo. Majoritariamente, as redes propostas foram treinadas em um notebook da Samsung, modelo np300e5m, versão 2017, com processador Intel Core i5-7200U de 7ª Geração (2,5 GHz até 3,1 GHz), 8 GB de RAM e um HD SSD de 512 GB de memória do tipo NVME m2 com taxas de 2400 MB/s e 1700 MB/s de leitura e escrita, respectivamente. Quando for apresentado o tempo decorrido para o treinamento de uma rede, pode-se assumir que esta fora treinada em um computador com tais especificações de hardware.

4.2 Banco de Dados e Pré-processamento dos sinais

Os sinais sonoros instrumentais utilizados do banco de dados "IRCAM Solo Instruments 2" [35] foram de contrabaixo, flauta, harpa e trompa, tocados na forma ordinária e podendo variar sua intensidade, a nota tocada e a corda (para o caso do contrabaixo). Como fora apresentado, estes sinais possuem períodos de silêncio que necessitam serem retirados para o treinamento dos classificadores.

Realizando o pré-processamento com a segmentação de silêncio de Zhang *et al.* [36] no sinal da Figura 3.3, tem-se como resultado o trecho em laranja do sinal apresentado na Figura 4.1.

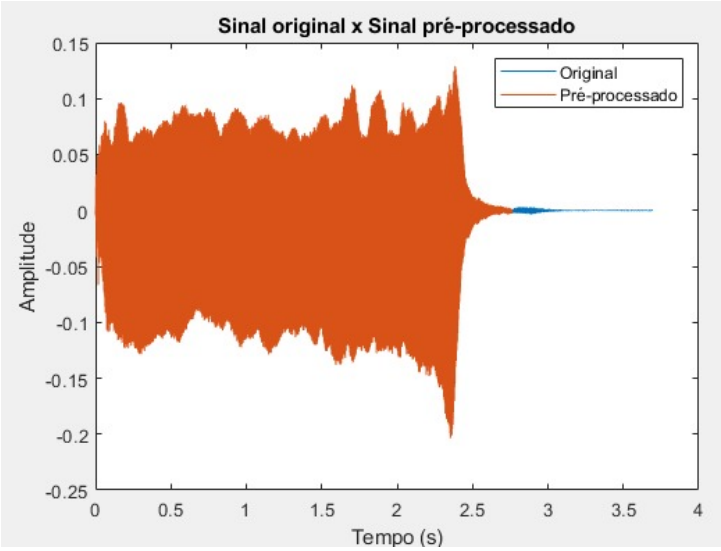


Figura 4.1: Resultado da segmentação de silêncio de Zhang *et al.* em um sinal de contrabaixo.

Observando o gráfico, percebe-se que a segmentação foi capaz de manter os trechos de maiores amplitudes do sinal e ainda manter o período de repouso do mesmo. Ao escutar o sinal de áudio, percebe-se que o mesmo soa até o silêncio, sem a presença de artefatos na finalização.

Aplicando o mesmo método para um sinal de harpa, temos o resultado da Figura 4.2.

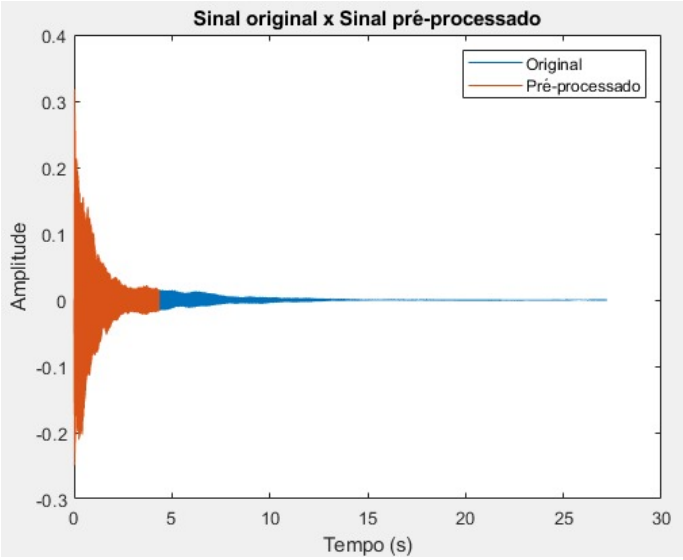


Figura 4.2: Resultado da segmentação de silêncio de Zhang *et al.* em um sinal de harpa.

Escutando o sinal cortado, percebe-se um artefato no final do áudio por este acabar de forma abrupta enquanto ainda era possível escutar a reverberação produzida pelo instrumento. Por conta desse problema, o sinal de harpa utilizado nos treinamentos fora cortado selecionando arbitrariamente e de forma manual em que ponto o sinal deve começar e terminar. Os pontos foram selecionados ao escutar o sinal de áudio e estimar o instante em que este passara a ser silêncio. O resultado do corte é visto na Figura 4.3.

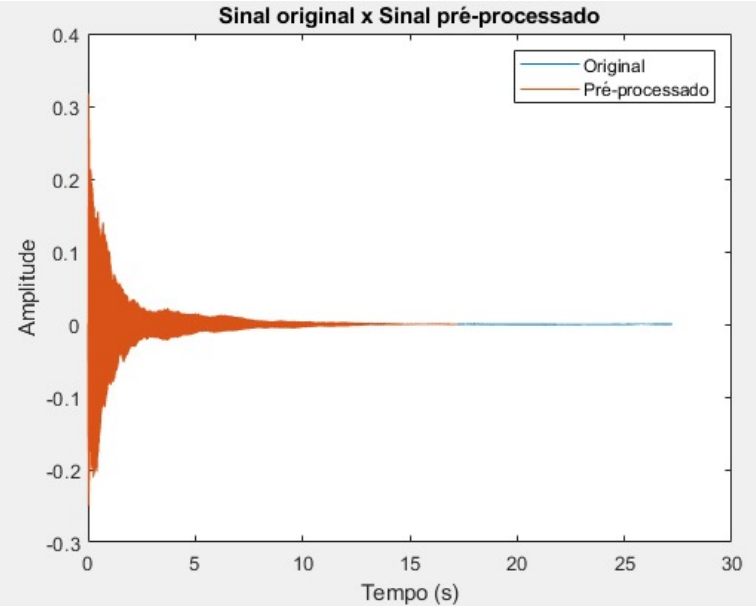


Figura 4.3: Resultado da remoção do silêncio pela com corte arbitrário em um sinal de harpa.

Neste corte, percebe-se que é mantida uma extensão muito maior do sinal. Ao escutar o mesmo, não é mais perceptível o artefato em seu final. O fato de a segmentação proposta por Zhang *et al.* realizar o corte em um instante de tempo prematuro pode ser atribuído à natureza da envoltória do sinal da harpa, com um ataque forte, sem sustentação e um repouso prolongado em baixas amplitudes devido à reverberação da corda do instrumento.

Os sinais de contrabaixo, flauta e trompa utilizados para treinamento das redes foram todos segmentados com o método de Zhang *et al.*, enquanto os sinais de harpa tiveram o silêncio segmentado por meio de um corte arbitrário.

4.3 Classificação baseada em descritores

Para realizar a classificação dos descritores de forma generalizada, foi utilizado o método de validação cruzada K-Fold com K igual a 10. Do conjunto separado para treinamento, 20% é utilizado para validação, totalizando em 72% do conjunto total utilizado para treino, 18% para validação e 10% para teste. Os testes foram feitos com sinais de apenas uma nota e com seis notas diferentes de cada instrumento, como é apresentado na Tabela 4.1.

Tabela 4.1: Configurações dos treinos de classificação.

Treino	**Rede**	**Neurônios**	**Notas**
CLASS1	MLP	10, 20, 30, 40	B4
CLASS2	MLP	10, 20, 30, 40	A4, B4, C4, D4, F4, G4
CLASS3	Elman	10, 20, 30, 40	B4
CLASS4	Elman	10, 20, 30, 40	A4, B4, C4, D4, F4, G4

Nesta tabela, a coluna "Treino" fornece um indicador para aquela configuração. "Rede" indica o tipo de rede neural treinada, podendo ser a rede MLP ou de Elman das Figuras 3.4 e 3.5, respectivamente. A quantidade de neurônios na camada intermediária é indicada na coluna "Neurônios". Os sinais de áudio utilizados são os de contrabaixo e flauta, sempre na mesma quantidade e tocando as mesmas notas. A coluna "Notas" indica quais notas musicais são tocadas em cada sinal de áudio[1]. Vale lembrar que cada sinal de áudio apresenta apenas uma nota sendo tocada, então esta coluna fornece também quantos sinais de cada instrumento estão sendo utilizados para o treinamento da rede. Todos os sinais utilizados são de contrabaixo e flauta. A Tabela 4.2 apresenta o resultado dos treinamentos de CLASS1 a CLASS4.

Tabela 4.2: Resultados dos treinos CLASS1 a CLASS4

Treino	Neurônios	Acurácia (%)		Épocas		Tempo
		Média	Desvio Padrão	Média	Desvio Padrão	
CLASS1	10	96,25	4,11	5156,60	3478,60	7m46s
	20	**97,19**	3,11	5663,00	3339,70	13m2s
	30	95,94	3,62	4445,60	2971,20	7m59s
	40	95,94	3,91	4580,70	3624,70	10m1s
CLASS2	10	90,97	1,97	5725,50	3199,40	10m52s
	20	91,48	1,87	6689,20	3559,80	22m49s
	30	90,97	1,75	7673,60	3229,00	40m16s
	40	91,16	2,61	6586,20	3632,00	48m57s
CLASS3	10	95,75	3,51	80,30	54,68	0m14s
	20	96,25	3,64	116,40	35,73	2m7s
	30	96,27	3,58	138,70	70,62	7m50s
	40	96,48	3,78	134,90	61,55	19m29s
CLASS4	10	95,43	1,88	279,10	139,96	4m8s
	20	96,11	1,40	160,20	93,96	11m57s
	30	95,55	1,50	142,00	40,85	30m50s
	40	**96,44**	1,55	177,30	52,06	102m47s

São fornecidos a média e o desvio padrão dos 10 treinamentos de K-Fold para cada quantidade de neurônios de cada configuração de treino. Além da acurácia, são apresentados a média de épocas e o desvio padrão para a rede chegar em seu estado final, e quanto tempo é necessário para o treinamento das 10 rodadas do K-Fold, em minutos e segundos. Em negrito são destacadas as melhores acurácias para a separação utilizando apenas um sinal com a nota B4 e para a separação utilizando as notas A4, B4, C4, D4, F4 e G4.

Pelas acurácias apresentadas nesta tabela, vemos que os descritores apresentados são suficientes para uma boa caracterização dos sinais utilizados, apresentando em todos os testes uma acurácia média superior a 90%. Curiosamente, a rede MLP obteve o melhor resultado que a

[1]Para referência das frequências das notas musicais, consulte a Tabela I.1, nos Anexos.

rede de Elman para a classificação em uma só nota. Era esperado que uma rede recorrente, que absorve informações da evolução do sinal ao longo do tempo, apresentaria um resultado melhor, considerando que informações da envoltória do sinal de um instrumento são importantes também para sua identificação. Porém, ao acrescentar mais sinais em notas diferentes (aumentando a complexidade do problema), percebe-se uma queda na acurácia da rede MLP, enquanto a da rede de Elman permanece alta, demonstrando a vantagem de uma rede recorrente para este tipo de problema.

É possível ainda fazer uma comparação com o sistema proposto por Anderson [8]: sua rede também fora uma MLP, classificando dois instrumentos diferentes com 6 notas de cada, mas utilizando 19 características do sinal para alcançar uma acurácia global de 75%, enquanto a rede proposta aqui nas mesmas condições alcançou uma acurácia média de até 91.48%. É importante ressaltar que os instrumentos e sinais usados para os testes são diferentes nos dois casos, e que enquanto Anderson classifica um sinal inteiro, a rede apresentada aqui classifica trechos do sinal. Apesar dessas diferenças, o grande incremento na acurácia mostra que um conjunto reduzido e bem escolhido de descritores já recupera informações importantes e suficientes do sinal, não havendo necessidade de todos as 19 características extraídas por Anderson.

Selecionando a rede com melhor classificação para a nota única (CLASS1 de 20 neurônios) e a com melhor classificação para as seis notas (CLASS4 de 40 neurônios), monta-se a Tabela 4.3, a qual apresenta a acurácia da classificação das redes treinadas em cada rodada de treino do K-Fold.

Tabela 4.3: Acurácia (%) de cada rodada do K-Fold dos treinos CLASS1 de 20 neurônios e CLASS4 de 40 neurônios

	Rodada de treino do K-Fold									
Treino	1	2	3	4	5	6	7	8	9	10
CLASS1	100	93,75	100	93,75	100	96,87	100	100	93,75	93,75
CLASS4	97,79	96,69	94,85	94,85	98,13	97,39	98,51	95,52	94,03	96,64

A matriz de confusão da rede com acurácia similar à media obtida para o caso do reconhecimento utilizando uma nota, CLASS1 com 20 neurônios rodada 6, é apresentada na Figura 4.4.

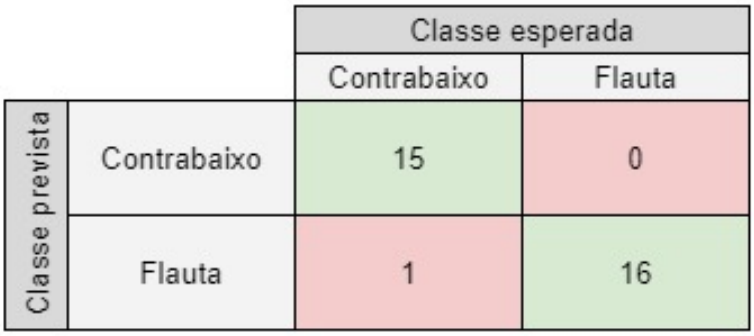


Figura 4.4: Matriz de confusão do treino CLASS1 com 20 neurônios no sexto treino K-Fold.

A matriz de confusão da rede com acurácia similar à melhor média obtida para o caso do reconhecimento utilizando várias notas, CLASS4 com 40 neurônios e rodada 10 é apresentada na Figura 4.5

		Classe esperada	
		Contrabaixo	Flauta
Classe prevista	Contrabaixo	157	6
	Flauta	3	102

Figura 4.5: Matriz de confusão do treino CLASS4 com 40 neurônios no sétimo treino K-Fold.

Observando as matrizes de confusão é visto outro fator que influencia na acurácia: para o caso com apenas uma nota (Figura 4.4), a quantidade testes realizados é relativamente pequena, com apenas uma falha fazendo com que a acurácia da classificação caia para 96.87% (nota-se que todos os treinamentos de uma nota com K-Fold alcançaram uma acurácia próxima a esse valor). Já no caso da classificação com mais notas (Figura 4.5), a quantidade de testes fora muito maior, conferindo maior confiança na capacidade de classificação das redes.

Além da acurácia, outra comparação entre a rede MLP e a rede de Elman digna de nota é o número de épocas e o tempo necessário para concluir o treinamento. Apesar de cada época de treinamento da rede MLP demorar menos tempo que as épocas da Elman, a rede Elman necessita de menor número de épocas (por volta de 50 vezes menos, em média) para finalizar o treinamento, apresentando um resultado comparável ou maior. Entretanto, percebe-se também que o aumento da quantidade de neurônios na camada intermediária afeta muito mais o tempo de treinamento da rede Elman do que da MLP.

Para demonstrar o funcionamento da melhor rede de Elman treinada, é apresentada na Figura 4.6 o gráfico de um sinal de contrabaixo e um sinal de flauta em sequência, junto com a classificação feita pela rede. O sinal azul se refere ao som do contrabaixo, tocado na nota A4, o azul ciano se refere ao som da flauta, tocada na nota C4, e em laranja é dada a predição da rede para cada trecho do sinal com quantidade de amostras suficiente para compor um vetor de descritores. Uma predição em 0 indica 100% de certeza da rede de que o trecho do sinal pertence à classe contrabaixo, e a predição em 1 indica 100% de certeza de que o trecho pertence à flauta.

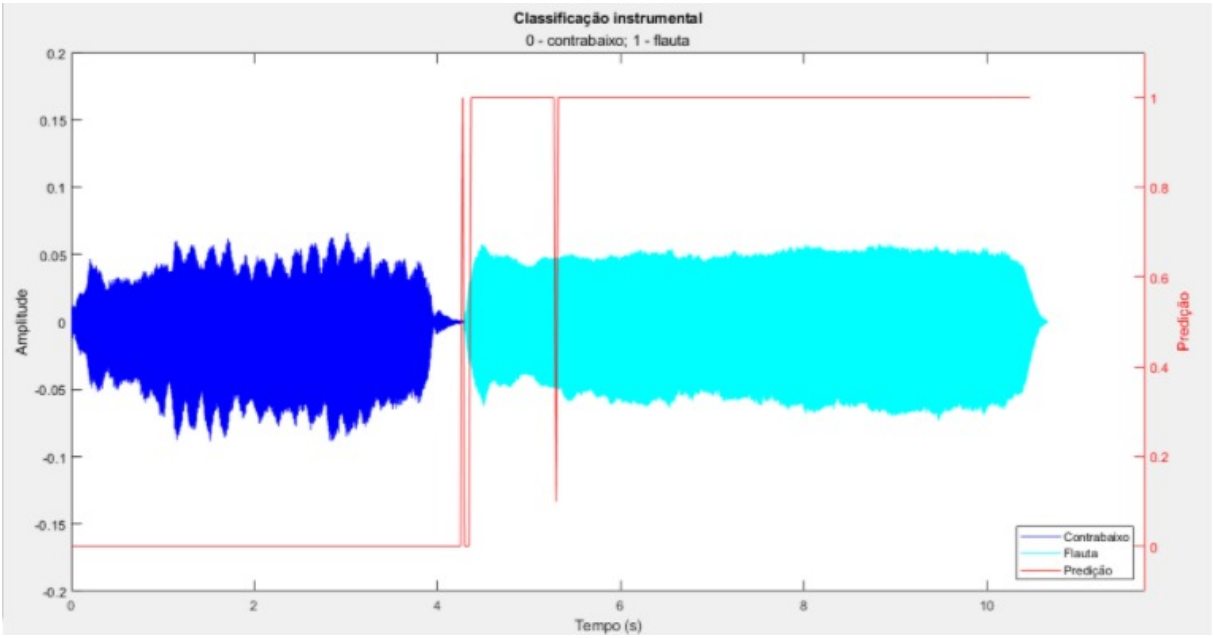


Figura 4.6: Exemplo de ondas artificiais usadas na separação.

Analisando o gráfico, pode-se ver que durante quase todo o sinal do contrabaixo a classificação fora feita corretamente, havendo apenas uma confusão ao final deste segmente, o que é compreensível pois o mesmo estava já próximo de se tornar silêncio. Começando o segmento da flauta, a rede rapidamente passa a classificar seus trechos como sendo de flauta, com uma confusão rápida em um pequeno trecho e classificando corretamente todo o resto de sua extensão. Este resultado mostra não só a robustez da rede em classificar corretamente um sinal composto por dois instrumentos, mas também sua capacidade em realizar a tarefa de forma online.

4.4 Separação na forma de onda

Enquanto o problema de classificação de fonte sonora requer apenas um resultado binário, classificando entre uma fonte ou outra, a separação requer uma precisão alta dos valores de saída para reconstruir com fidelidade as fontes que originaram a mistura, o que confere uma complexidade consideravelmente maior ao problema. Com isto em mente, os testes realizados na separação foram divididos em diferentes graus de complexidade, passando pelo caso simples da separação de sinais artificiais de forma especializada e generalizada, e incrementando a complexidade ao treinar a rede para separação de sinais instrumentais.

4.4.1 Sinais artificiais

Os sinais artificiais utilizados foram ondas senoidais, quadradas e triangulares, com frequência de amostragem de 44,1 kHz (igual aos áudios das gravações instrumentais). Um exemplo de cada uma dessas fontes pode ser visto na Figura 4.7.

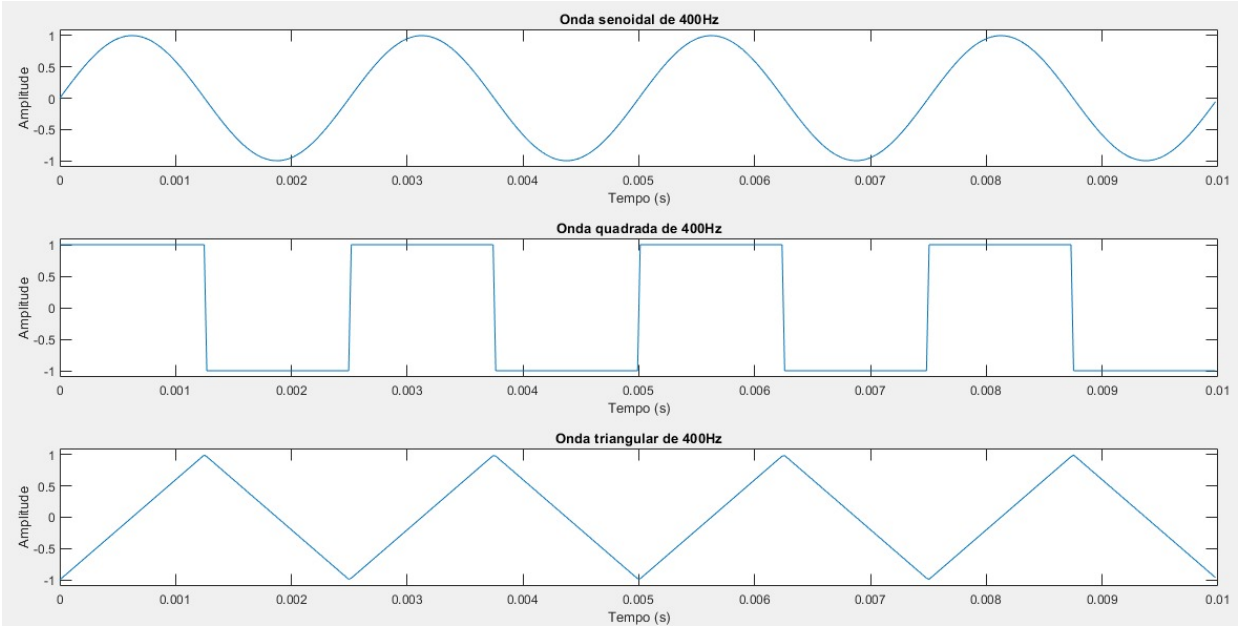


Figura 4.7: Exemplo de ondas artificiais usadas na separação.

A frequência das ondas se mantém constante ao longo da duração do sinal, mas cada sinal pode ter uma frequência diferente. Todos os sinais possuem a mesma fase e amplitude. A rede recebe um sinal com a mistura entre uma onda senoidal, uma quadrada e uma triangular (formada com a fórmula da Equação 3.2), e deverá retornar três saídas sequenciais contendo cada uma dessas ondas.

As redes desenvolvidas são as apresentadas nas figuras 3.6 e 3.7, com 3 neurônios na camada de saída (um para cada um dos sinais artificiais usados como fonte). A função de loss, utilizada para mensurar a diferença entre a saída obtida e a saída esperada, é dada pela metade do Erro Quadrático Médio (MSE, Mean Squared Error)[2], como apresentado pela Equação 4.1:

$$loss = \frac{1}{2 \cdot S} \cdot \sum_{i=1}^{S} \sum_{j=1}^{N_F} (f_{ij} - f_{eij})^2 \tag{4.1}$$

onde S é o número de amostras, N_F é o número de fontes,f_{ij} é a amostra j da fonte real i e f_{eij} é a amostra j da fonte estimada i.

A quantidade de épocas em que a rede é treinada pode ser determinada de três formas: por um valor fixo escolhido em momento anterior ou por melhoria do loss ou do SDR. Esse último é feito da seguinte forma: a rede é treinada em uma quantidade pequena de épocas e testada ao final, calculando seu SDR médio. O processo é repetido diversas vezes, treinando a rede em seu último estado e marcando a com maior SDR médio até o momento como a melhor rede. Para que tal rede seja tratada como a rede final, devem ser treinadas mais uma determinada quantidade de épocas sem que o SDR médio aumente. A melhoria do loss segue um processo semelhante, considerando a rede melhor como aquela cujo loss é o menor.

[2] `https://www.mathworks.com/help/deeplearning/ref/regressionlayer.html`

Foram feitos testes tanto com treino especializado, aplicando apenas uma única mistura durante todo o treinamento, quanto com treino generalizado, aplicando diversas combinações dos sinais artificiais.

A Tabela 4.4 define as configurações de cada experimento de treino feito com sinais artificiais, de forma especializada (ART1 a ART9) e generalizada (ART10).

Tabela 4.4: Configurações dos treinos com ondas artificiais.

Treino	**Rede**	**Neurônios**	**Parada**	**Frequências (Hz)**	**Duração (s)**
ART1	LSTM	40, 70, 100, 130, 160	SDR	(400,400,400)	0.01
ART2	LSTM	40, 70, 100, 130, 160	SDR	(200,300,500)	0.01
ART3	LSTM	40, 70, 100, 130, 160	SDR	(200,600,1000)	0.01
ART4	LSTM	40, 70, 100, 130, 160	SDR	(1000,200,600)	0.01
ART5	LSTM	40, 70, 100, 130, 160, 190, 220, 250	loss	(400,400,400)	1
ART6	LSTM	40, 150	loss	(440,440,440)	1
ART7	LSTM	40	loss	(200,300,500)	1
ART8	Elman	40	epochs	(440,440,440)	0.01
ART9	Elman	40	epochs	(440,440,440)	1
ART10	LSTM	40, 70, 100, 130, 160, 500, 1000	SDR	Generalizado	0.01

A coluna “Treino” fornece um indicador para o treino com tais configurações, usado nas tabelas seguintes, com “ART” indicando o treino com sinal artificial. A coluna “Rede” indica qual arquitetura de rede neural fora utilizada, podendo ser a LSTM ou a Elman. A coluna “Neurônios” se refere à quantidade de neurônios (ou de células de memória, no caso da LSTM) presente na camada intermediária da rede naquela configuração. A coluna “Parada” indica se a parada do treinamento fora feita devido ao SDR testado não aumentar mais, ao loss não diminuir mais ou se chegou a um número de épocas determinado. A coluna “Frequências (Hz)” se refere à frequência das ondas de entrada, no formato (f1,f2,f3), com f1 sendo a frequência da onda senoidal, f2 a frequência da onda quadrada e f3 a da onda triangular. A mistura das componentes é realizada pela Equação 3.2 com os sinais na frequência apresentada nesta coluna. Por fim, a coluna “Duração (s)” indica a duração em segundos dos sinais utilizados no treinamento

Os resultados de ART1 a ART9, referentes aos treinos especializados com mistura de sinais senoidal, quadrado e triangular, são apresentados na Tabela 4.5.

Tabela 4.5: Resultado dos treinos ART1 a ART9,

Treino	Neurônios	SDR				Loss
		Senoide	Quadrada	Triangular	Média das Fontes	
ART1	40	48,1812	50,4769	46,5406	48,40	9,8000e-6
	70	47,8408	52,2041	49,3355	49,79	1,0430e-5
	100	47,9366	54,4632	49,9463	50,78	1,0640e-5
	130	50,3431	54,6523	48,1439	**51,05**	**8,1700e-6**
	160	45,7671	51,1363	46,8881	47,93	1,6990e-5
ART2	40	49,7910	46,5943	41,5461	**45,98**	**9,4000e-6**
	70	45,4718	46,9394	40,8384	44,42	1,0770e-5
	100	45,6333	44,9229	39,4264	43,33	1,2980e-5
	130	46,7888	44,5000	39,2000	43,50	1,2780e-5
	160	45,1796	46,2118	40,5117	43,97	1,2520e-5
ART3	40	47,5417	43,6548	41,0639	44,09	1,3560e-5
	70	46,7446	44,2824	38,8487	43,29	1,7870e-5
	100	47,0037	45,626	40,254	**44,29**	**1,1900e-5**
	130	45,8276	43,118	40,588	43,18	2,0610e-5
	160	47,7685	44,8638	38,4083	43,68	1,9360e-5
ART4	40	41,6238	49,2233	40,7942	**43,88**	1,2630e-5
	70	43,5262	47,1485	39,0934	43,26	1,5070e-5
	100	42,1569	46,8723	41,7815	43,60	**1,1390e-5**
	130	40,594	49,6986	40,6322	43,64	1,6790e-5
	160	42,6368	45,3145	40,3114	42,75	1,4970e-5
ART5	40	27,752	27,5403	21,9873	25,76	7,0000e-4
	70	25,2509	23,765	19,3371	22,78	1,5000e-3
	100	30,8762	22,8351	21,1304	24,95	1,3000e-3
	130	24,6045	24,4202	21,4859	23,50	1,0000e-3
	160	28,6988	25,2093	25,6943	26,53	7,0000e-4
	190	27,7168	24,4424	24,0467	25,40	1,0000e-3
	220	32,4374	26,2057	25,7241	**28,12**	**4,0000e-4**
	250	25,9515	24,3163	21,4025	23,89	1,2000e-3
ART6	40	30,2339	28,5924	27,6233	28,82	1,8779e-4
	150	32,9723	31,1142	25,293	**29,79**	**1,3175e-4**
ART7	40	7,8505	18,5029	8,6652	**11,67**	**1,0100e-2**
ART8	40	12,1054	62,7843	17,7329	**30,87**	**4,3000e-3**
ART9	40	9,6747	34,6979	25,0964	**23,16**	**4,5000e-3**

Como o treino é especializado em uma única mistura de entrada, o SDR é calculado com a saída da rede treinada nessa mesma mistura, utilizando a função fornecida por Vincent *et al.*[3]. A qualidade da separação é avaliada para cada uma das saídas, sendo fornecida também a média do SDR das saídas. Para comparação, também é apresentado o Loss da rede, calculado com a Equação 4.1. Em negrito estão os melhores resultados para cada configuração de treino, em relação à quantidade de neurônios utilizada.

Apresentando os resultados do treinamento especializado com maior SDR médio (ART1 com 130 neurônios), a Figura 4.8 mostra as saídas esperadas e obtidas de cada canal na separação.

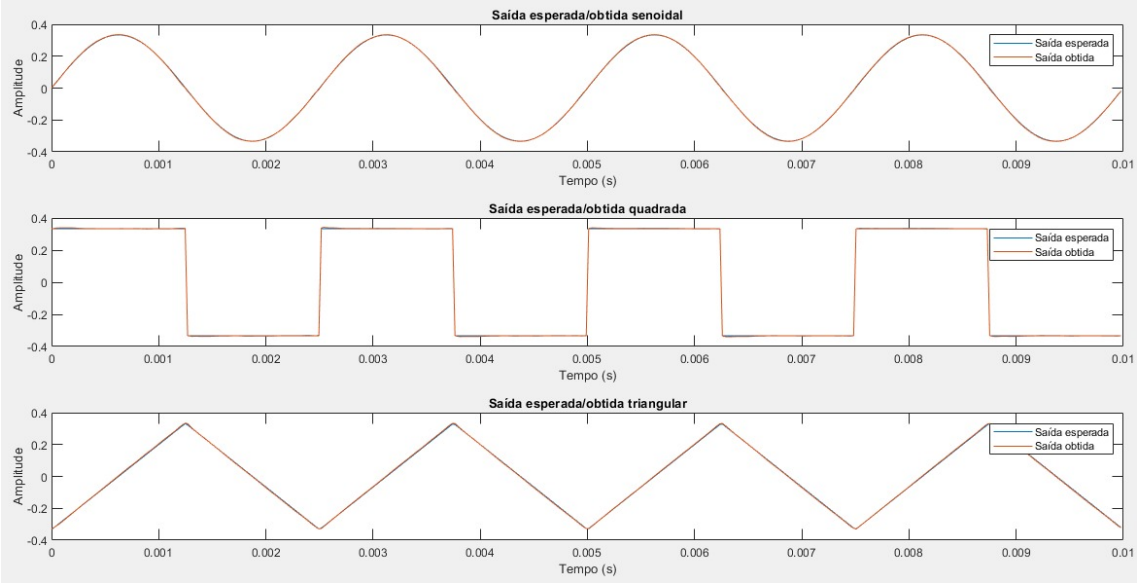


Figura 4.8: Saídas esperadas e obtidas do teste do ART1 com 130 neurônios.

Observando tais formas de onda, pode-se afirmar que a separação fora bem feita, com cada sinal de saída acompanhando muito bem o sinal esperado. Para uma análise mais profunda dessa separação, são apresentados também os espectrogramas dos sinais deste teste nas figuras 4.9, 4.10 e 4.11.

[3]Disponível em `https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/48623/versions/1/previews/SOLO_demo/bss_eval_sources.m/index.html`

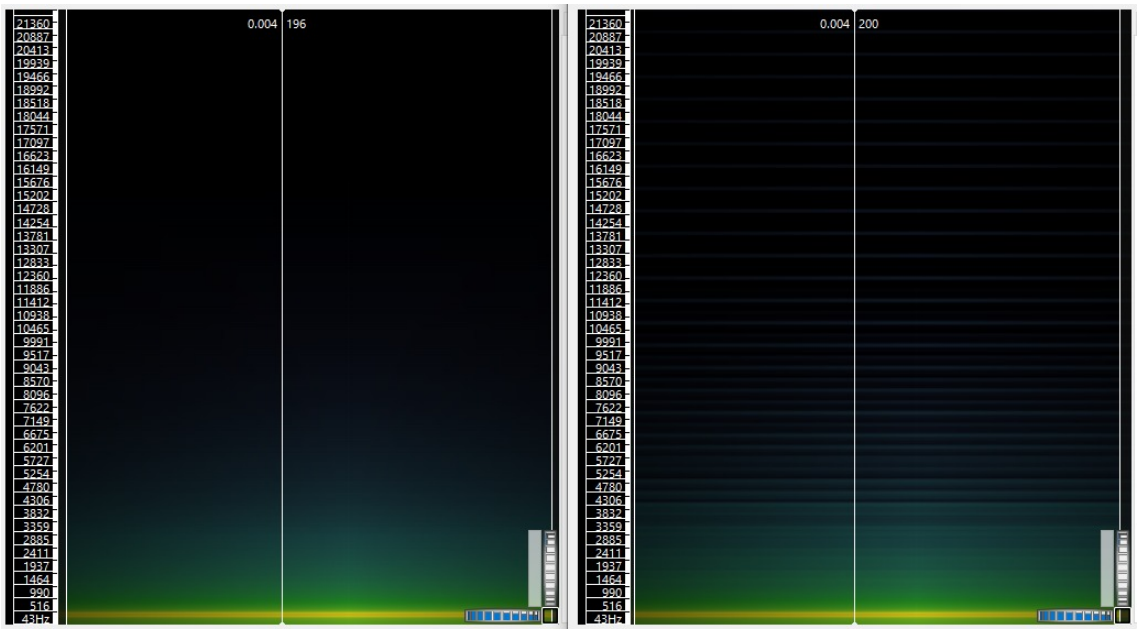


Figura 4.9: Espectrogramas das saídas esperada e obtida senoidais do ART1 de 130 neurônios.

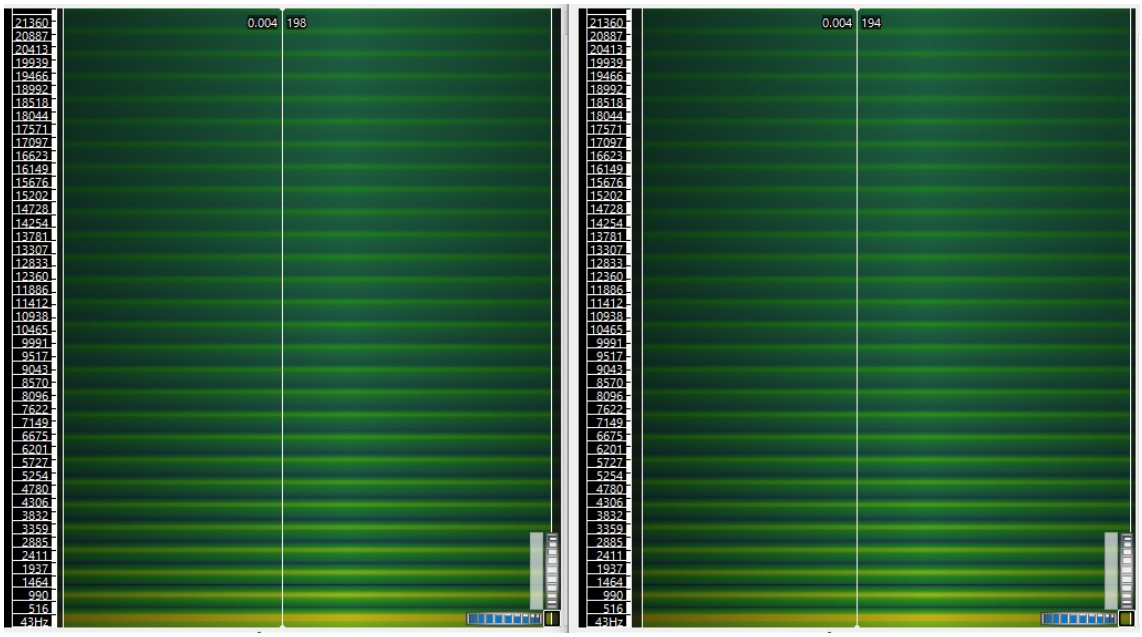


Figura 4.10: Espectrogramas das saídas esperada e obtida quadradas do ART1 de 130 neurônios.

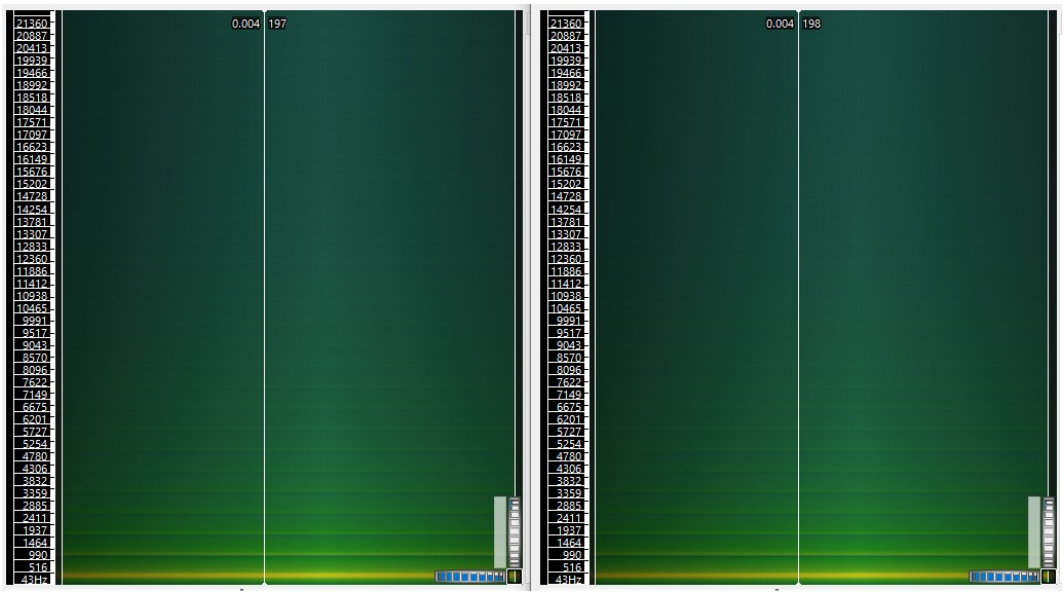


Figura 4.11: Espectrogramas das saídas esperada e obtida triangulares do ART1 de 130 neurônios.

Em cada uma das figuras, o espectrograma da esquerda corresponde ao sinal de saída esperado e o da direita, ao obtido, para fins de comparação. Novamente, nota-se que os sinais separados estão muito próximos das saídas esperadas, sendo difícil perceber alguma diferença entre as elas para as ondas quadrada e triangular. Ao escutar os áudios, percebe-se também que são extremamente semelhantes. Já pela saída senoidal, conforme mostrado no espectrograma da direita da Figura 4.9, percebe-se um leve acréscimo de frequências superiores, o que é quase imperceptível ao escutar o áudio e aparece como um ruído de alta frequência e baixíssima amplitude. Como uma onda senoidal é um som puro e sem harmônicos, o acréscimo de qualquer ruído é perceptível e o sinal deixa de ser uma senoide pura.

A Figura 4.12 mostra o resultado com maior SDR em que as ondas de entrada possuem frequências diferentes.

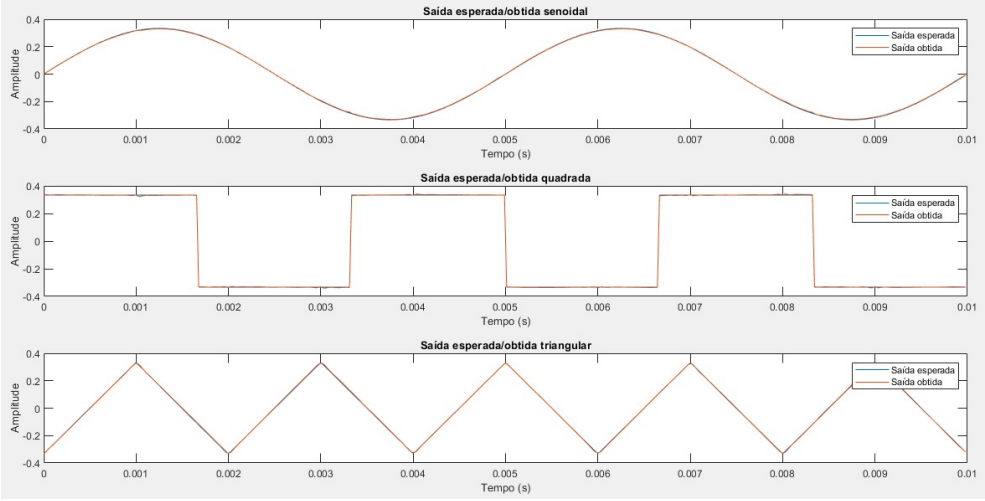


Figura 4.12: Saídas esperadas e obtidas do teste do ART2 com 40 neurônios.

Como pode ser visto, mesmo que as fontes possuam frequências diferentes, o sistema é capaz de fornecer uma boa separação, e seu tempo de resposta é pequeno, com o sinal de saída acompanhando o esperado desde as primeiras amostras. Destaca-se que tanto ART1 quanto ART2 utilizam a melhoria do SDR como critério de parada. A fim de comparar tal critério com o da melhoria do Loss, apresenta-se na Figura 4.13 as saídas do treino ART6 com 150 neurônios, que apresenta o menor Loss dentre os treinos com tal critério.

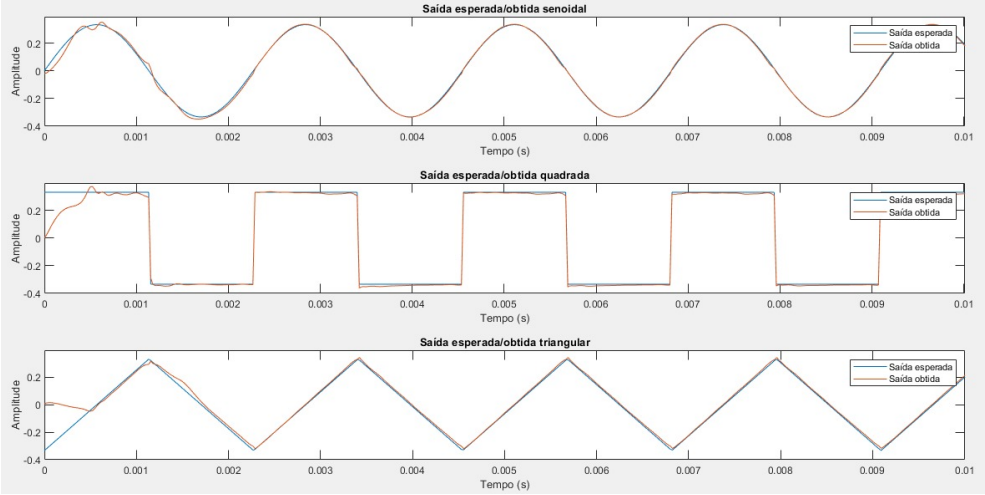


Figura 4.13: Saídas esperadas e obtidas do teste do ART6 com 150 neurônios.

No escopo geral, pode-se afirmar que a separação fora bem feita, com o sinal obtido acompanhando o esperado. Entretanto, o tempo de resposta é maior que o dos dois resultados já apresentados, com um erro maior na separação nas primeiras amostras. Tal fato indica uma vantagem do critério de melhoria do SDR sobre a melhoria do Loss.

Todos os resultados apresentados até agora foram obtidos com o treinamento da rede LSTM. Para comparação com a rede Elman, é apresentado na Figura 4.14 o resultado do melhor treinamento utilizando tal rede.

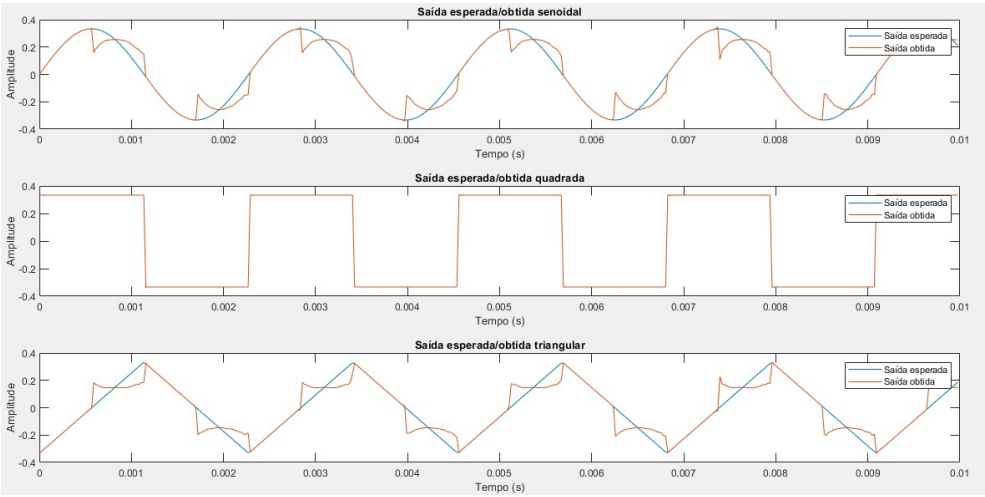


Figura 4.14: Saídas esperadas e obtidas do teste do ART8 com 40 neurônios.

Esta rede não retornou uma separação aceitável das ondas senoidal e triangular, mas possui a melhor separação da onda quadrada, tanto pelo valor de SDR quanto visualmente. Uma suposição que pode ser feita com esse resultado é que a rede é boa para sinais altamente discretizados (a onda quadrada pode apresentar apenas dois valores distintos).

No caso do treino generalizado, cada onda senoidal, quadrada e triangular original pode possuir frequência entre 200 e 1000 Hz, variando em 100 Hz entre cada opção, totalizando 9 possibilidades de sinais para cada fonte. Os sinais possuem todos 0.01 segundos de duração. As entradas do sistema são formadas por cada uma das possíveis combinações entre as fontes previamente separadas, formando um total de $9 \cdot 9 \cdot 9 = 729$ misturas de entrada.

O treino generalizado é feito com K-Fold, com $K = 5$ para cada treino, e utilizando como critério de parada a melhoria do SDR médio, com uma modificação: como o processo de calcular o SDR médio de todas as entradas de treino (aproximadamente 583 entradas) é demorado, este cálculo é feito com apenas 10% das entradas durante o treinamento. O SDR atribuído à rede treinada é calculado com todas as entradas separadas para teste.

Os resultados do treino feito de forma generalizada, ART10, são apresentados na Tabela 4.6.

Tabela 4.6: Resultado dos treinos ART10

Neurônios	SDR Médio			
	Senoidal	Quadrada	Triangular	Média entre fontes
40	4,2540	6,9564	3,2519	4,8208
70	4,1047	7,0255	3,3407	4,8236
100	3,9966	7,0820	3,3652	4,8146
130	4,0303	7,1134	3,2015	4,7817
160	4,0077	7,1584	3,3502	4,8388
500	**4,1720**	**7,1645**	**3,5707**	**4,9691**
1000	4,0941	7,1676	3,4142	4,892

A cada rodada de treinamento do K-Fold, é calculado o SDR da separação de aproximadamente 146 misturas selecionadas unicamente para teste, e então obtida o SDR médio de cada saída. A Tabela 4.6 mostra o resultado final do K-Fold (média do SDR médio de cada saída para cada rodada de treinamento) para as diferentes quantidades de neurônios na camada intermediária. Além disso, apresenta também a média do SDR entre as fontes. Em negrito está marcada a rede com média total mais alta.

Nota-se que os valores de SDR são bem menores que os obtidos na separação especializada, apresentando pouca variação ao alterar a quantidade de neurônios. Para analisar alguma das saídas obtidas, é selecionanda a rede com melhor resultado (a de 500 neurônios), apresentando na Tabela 4.7 alguns dados estatísticos do SDR calculado com as misturas de teste para cada uma das rodadas de treinamento, sendo eles: média, mediana, maior e menor SDR. Estes são dados para cada uma das saídas, e uma média dentre elas é fornecida também.

Tabela 4.7: Estatísticas de cada grupo k do K-Fold do treino ART10 com 500 neurônios

Rodada	Estatísticas	SDR Médio			
		Senoide	Quadrada	Triangular	Média entre fontes
1	Média	4,2519	7,4128	3,5805	5,0817
	Mediana	3,1	6,1174	2,7752	3,9975
	Melhor	16,1414	25,7838	24,3871	22,1041
	Pior	3,1702	3,5904	2,1339	2,9648
2	Média	4,3333	7,1384	3,6757	5,0491
	Mediana	3,1845	6,1294	3,0495	4,1211
	Melhor	12,0154	22,6231	19,5238	18,0541
	Pior	2,9813	3,4168	1,6439	2,6807
3	Média	4,2388	7,2654	3,541	5,0151
	Mediana	3,0028	6,1099	2,6329	3,9152
	Melhor	18,7838	25,3029	25,8412	23,3093
	Pior	3,0657	2,9407	2,6464	2,8843
4	Média	4,2078	7,2128	3,543	4,9879
	Mediana	2,9955	6,0666	2,7434	3,9352
	Melhor	16,916	24,2092	25,363	22,1627
	Pior	3,3649	3,0215	1,4723	2,6196
5	Média	3,8281	6,793	3,5134	4,7115
	Mediana	3,1653	6,0588	3,0843	4,1028
	Melhor	12,3088	24,7302	17,7841	18,2744
	Pior	3,5605	1,6911	2,7348	2,6621
Total	Média	4,1720	7,1645	3,5707	4,9691

Um ponto a se destacar da tabela é o quanto o SDR em cada treino pode variar, indo de valores próximos a 3 no pior caso até valores acima de 20 em alguns dos treinos. Também percebe-se que, em média, a onda quadrada obtida apresenta um SDR melhor que as ondas senoidal e triangular. Possivelmente a natureza binária da onda facilita sua separação, ou talvez a presença de harmônicos superiores com alta amplitude (mais altos que na onda triangular) mascare a infiltração de outros sinais.

Dentre os testes realizados no ART10 com 500 neurônios, é apresentado um teste representativo feito na segunda rodada com uma mistura de uma onda senoidal de 500Hz, uma quadrada de 800Hz e uma triangular de 500Hz, cuja separação apresenta um SDR médio das fontes de 4,9656. As formas de onda das saídas esperadas e obtidas deste teste podem ser vistas na Figura 4.15 e os espectrogramas comparando cada uma das saídas obtidas com as esperadas nas Figuras 4.16, 4.17 e 4.18.

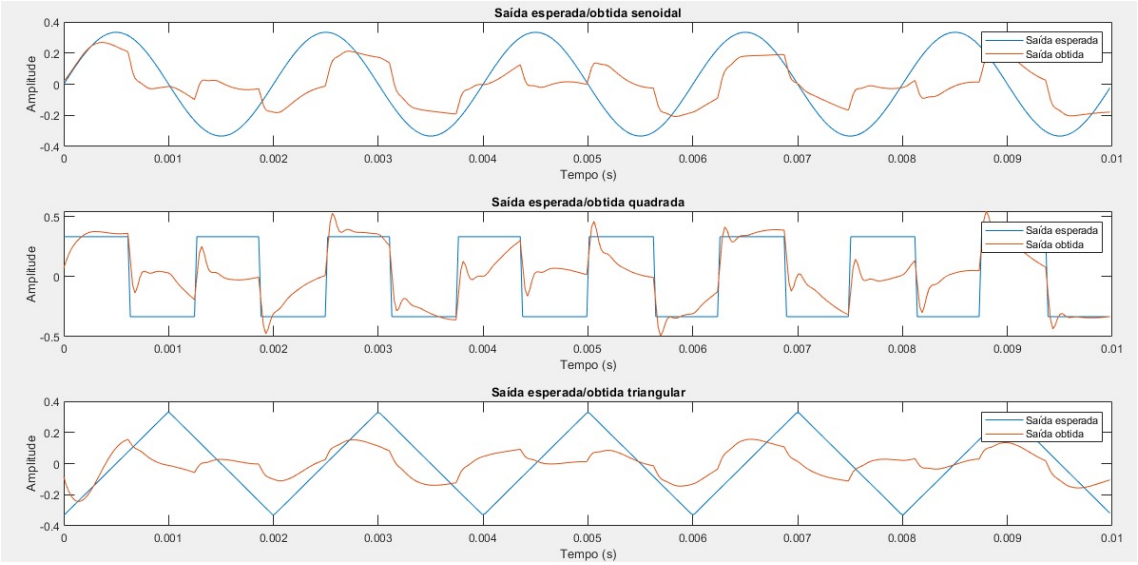


Figura 4.15: Saída esperada e obtida de um teste mediano feito em ART10 com 500 neurônios.

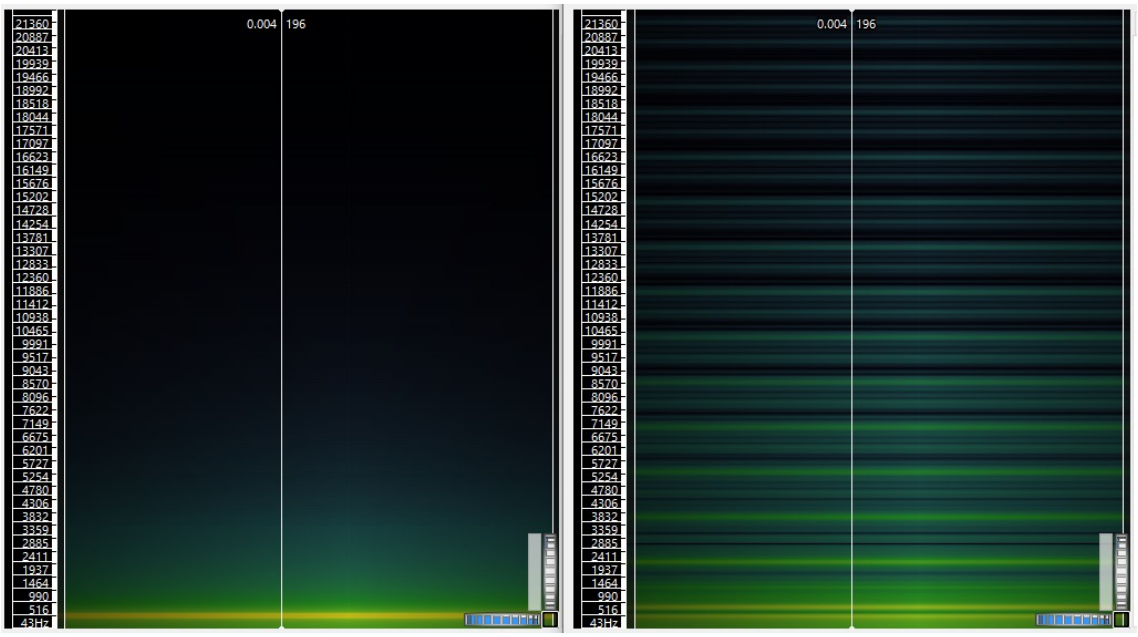


Figura 4.16: Espectrogramas dos sinais senoidais da Figura 4.15 (esperada na esquerda, obtida na direita).

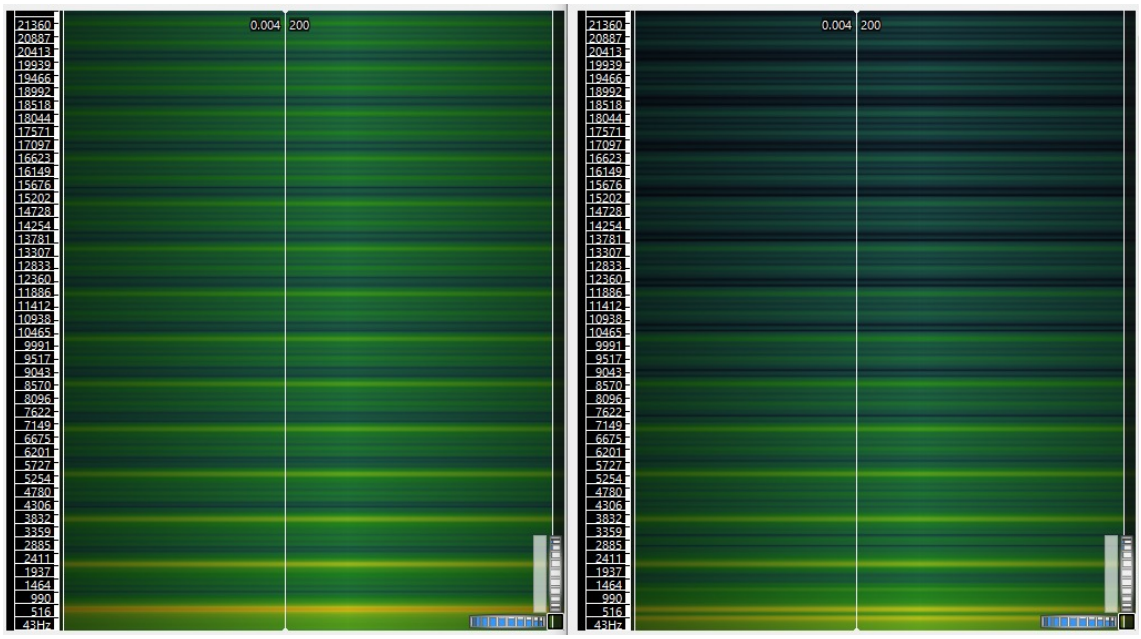


Figura 4.17: Espectrogramas dos sinais de onda quadrada da Figura 4.15 (esperada na esquerda, obtida na direita).

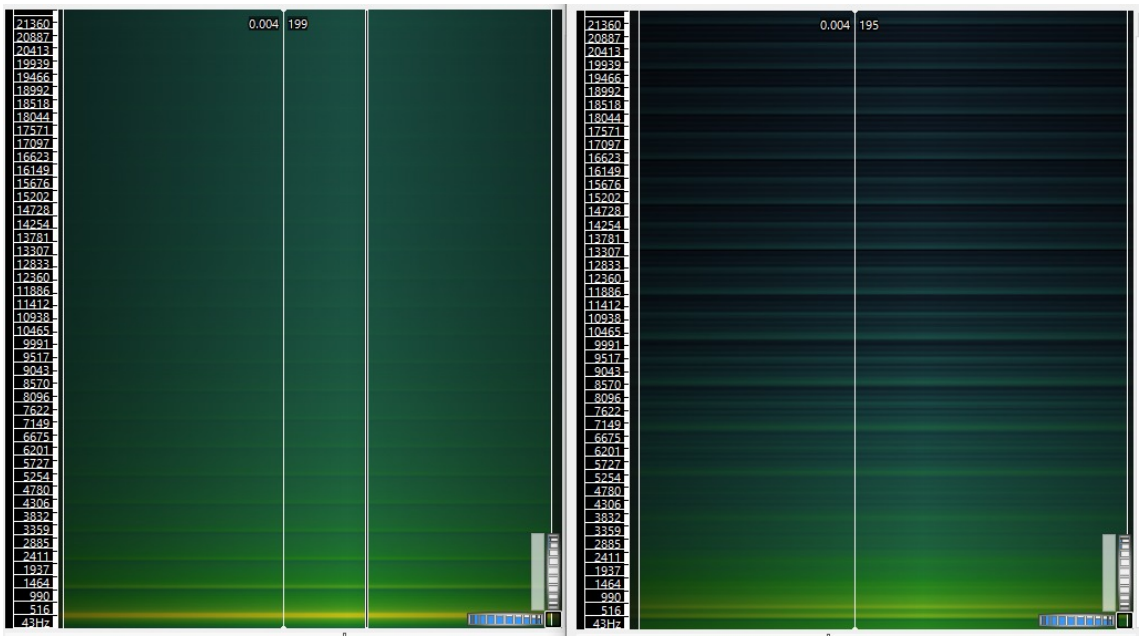


Figura 4.18: Espectrogramas dos sinais de onda triangular da Figura 4.15 (esperada na esquerda, obtida na direita).

Analisando as figuras, percebe-se que a separação não teve êxito. Um dos pontos a se notar é que as saídas senoidal e triangular ficaram muito semelhantes entre si, com a onda quadrada chegando o mais próximo do esperado, mas com muitas distorções ainda. As fontes possuem frequências diferentes, e pelos espectrogramas pode-se ver que ambas estão presentes em todos os sinais, com a senoide sofrendo de interferência de diversas outras frequências, pertencentes aos

harmônicos das outras ondas. Quanto às ondas quadrada e triangular, nota-se também que seus harmônicos superiores possuem amplitude menor do que o esperado.

Para analisar um resultado com mais chance de apresentar boa separação, é escolhido aquele com maior SDR médio, feito na terceira rodada com uma mistura de ondas senoidal, quadrada e triangular em 800 Hz cada, cuja separação apresenta um SDR médio das fontes de 23,3093. As formas de onda das saídas esperadas e obtidas deste teste podem ser vistas na Figura 4.19, e os espectrogramas comparando cada uma das saídas obtidas com as esperadas nas Figuras 4.20, 4.21 e 4.22.

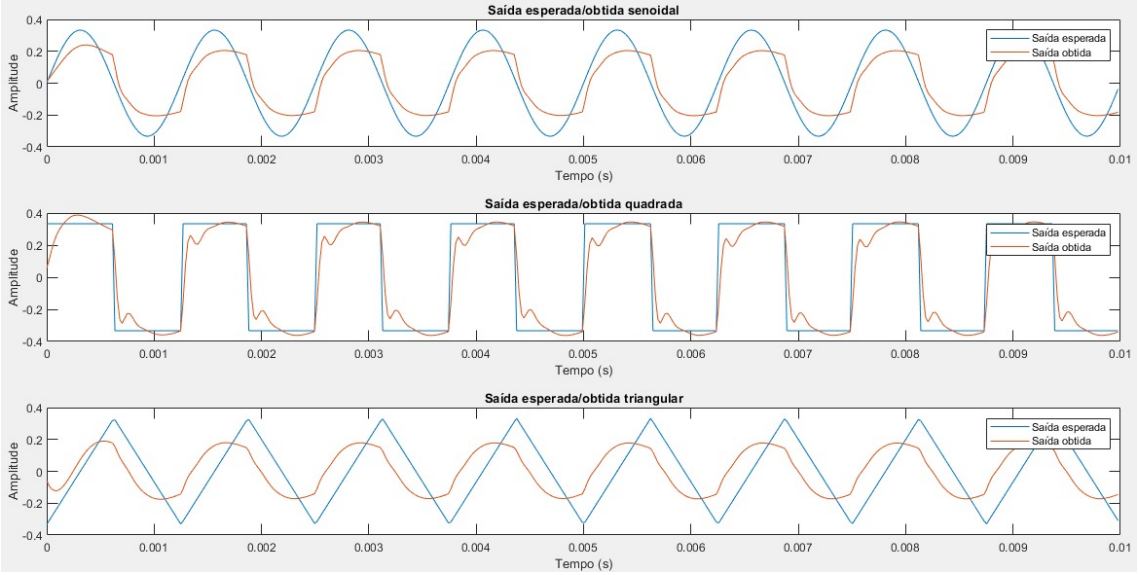


Figura 4.19: Saída esperada e obtida do melhor teste feito em ART10 com 500 neurônios.

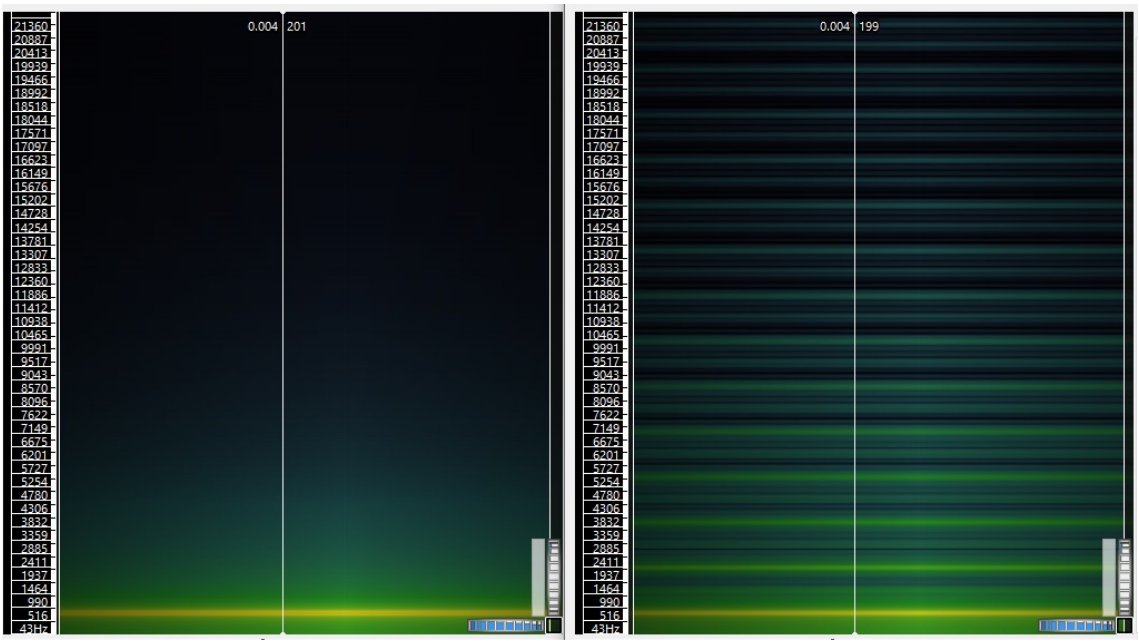


Figura 4.20: Espectrogramas dos sinais senoidais da Figura 4.19 (esperada na esquerda, obtida na direita).

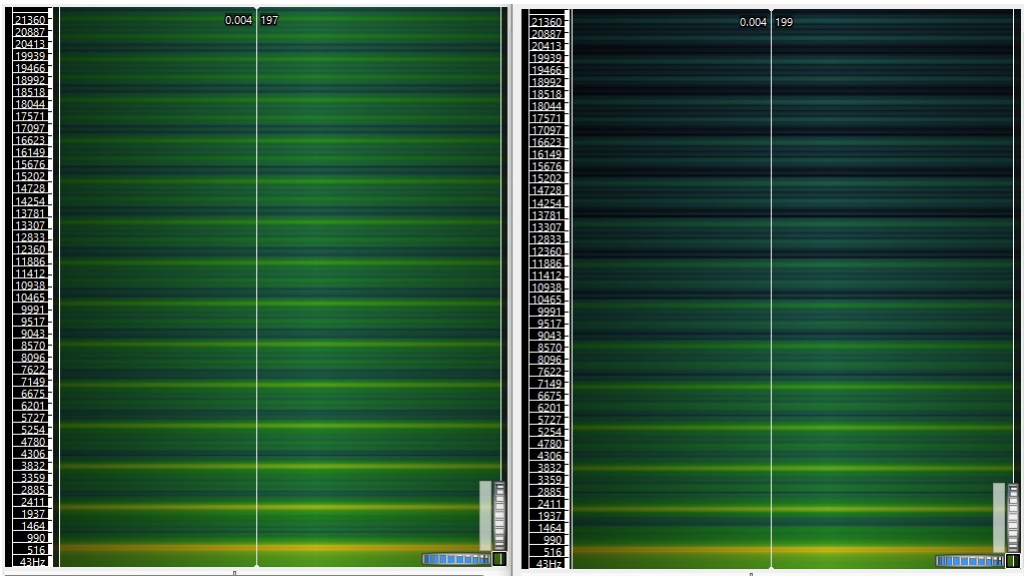


Figura 4.21: Espectrogramas dos sinais de onda quadrada da Figura 4.19 (esperada na esquerda, obtida na direita).

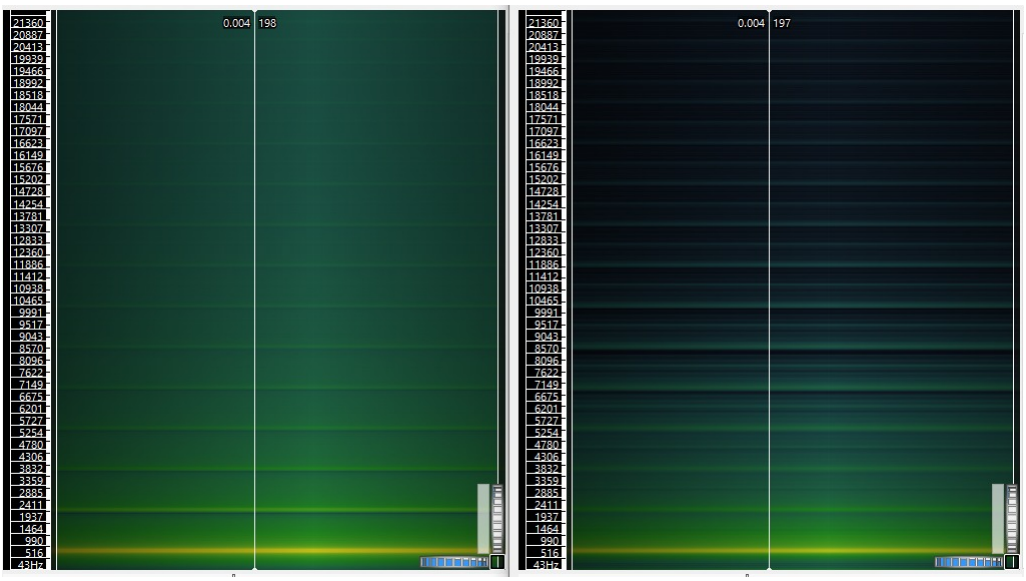


Figura 4.22: Espectrogramas dos sinais de onad triangular da Figura 4.19 (esperada na esquerda, obtida na direita).

Neste resultado, percebe-se novamente alguns pontos levantados para o resultado representativo anterior: as saídas senoidal e triangular ficaram muito semelhantes, com a quadrada chegando mais próximo do esperado; e seus espectrogramas mostram que muitas frequências foram inseridas no sinal senoidal, correspondentes aos harmônicos superiores das ondas triangular e quadrada, e as amplitudes dos harmônicos superiores destes ficaram reduzidas em relação ao valor esperado. Como todos os sinais estão na mesma frequência e as ondas quadrada e triangular possuem os mesmos harmônicos (com amplitudes diferentes), torna-se difícil identificar a interferência de um sinal no outro.

4.4.2 Sinais de instrumentos

Na etapa de separação de sinais instrumentais, a rede receberá uma mistura contendo dois sinais de áudio, cada um de um instrumento diferente, misturados pela fórmula da Equação 3.2. A rede deverá retornar cada um dos originais em canais diferentes. Nesta etapa, são treinadas apenas redes especializadas, com o intuito de analisar se tais redes conseguem aprender um caso simples, mas já com a complexidade de sinais instrumentais.

As redes poderão ser a de Elman ou a LSTM apresentadas nas Figuras 3.6 e 3.7, respectivamente, podendo variar a quantidade de neurônios (ou células de memória) na camada intermediária em ambos casos. A quantidade de épocas de treinamento pode ser fixa ou utilizando a melhoria do loss ou do SDR.

Quanto aos sinais utilizados, estes podem ser de uma mistura de contrabaixo e flauta, por serem sons semelhantes, ou de harpa e trompa, por serem sons diferentes. O treinamento pode ser feito com apenas um pedaço de cada sinal de áudio ou com o sinal inteiro. Um resumo dos treinos feitos é apresentado na Tabela 4.8.

Tabela 4.8: Configurações dos treinos com sinais instrumentais.

Treino	**Rede**	**Neurônios**	**Parada**	**Instrumentos**	**Notas**	**Trecho**
INS1	LSTM	5, 100, 500	loss	(cb,fl)	(E4,E4)	Inteiro
INS2	LSTM	5, 100, 500	loss	(cb,fl)	(E4,E4)	0.28 s
INS3	LSTM	2, 12, 22, 32, 42, 52	epoch	(cb,fl)	(E4,E4)	0.28 s
INS4	LSTM	10, 15, 20, 25, 30, 35, 40, 45, 50	loss	(hp,tp)	(A#1,G4)	Inteiro
INS5	Elman	14	epoch	(cb,fl)	(E4,E4)	0.28 s
INS6	Elman	14	epoch	(cb,fl)	(E4,E4)	inteiro
INS7	Elman	14, 40	epoch	(hp,tp)	(A#1,G4)	inteiro
INS8	LSTM	20, 60, 80, 100, 200	SDR	(hp,tp)	(A#1,G4)	3 s

A coluna “Treino” fornece um indicador para aquela configuração de treino, a ser utilizada ao apresentar os resultados. A coluna “Rede” indica a rede utilizada para treino, podendo ser LSTM ou Elman, com a quantidade de neurônios (ou células de memória) na camada intermediária especificada na coluna “Neurônios”. A coluna “Parada” indica qual método utilizado para determinar a quantidade de épocas treinadas, podendo ser um valor determinado de épocas ou por melhoria do Loss ou do SDR. A coluna “Instrumentos” se refere a quais instrumentos geraram os sinais de áudio que são misturados para servir como entrada da rede, no formato (i1,i2), sendo i1 o primeiro instrumento e i2 o segundo. Os instrumentos podem ser cb para contrabaixo, fl para flauta, hp para harpa e tp para trompa. A coluna “Notas” indica qual nota musical cada instrumento está

tocando[4], no formato (n1,n2), com n1 e n2 se referindo às notas dos instrumentos i1 e i2. Por fim, a coluna "Trecho" indica se a nota fora treinada com o sinal de áudio inteiro ou com um pequeno trecho do mesmo, sendo indicada a duração do trecho neste caso.

O resultado dos treinos nas configurações da Tabela 4.8 é apresentado na Tabela 4.9.

Tabela 4.9: Resultados dos treinos INS1 a INS8

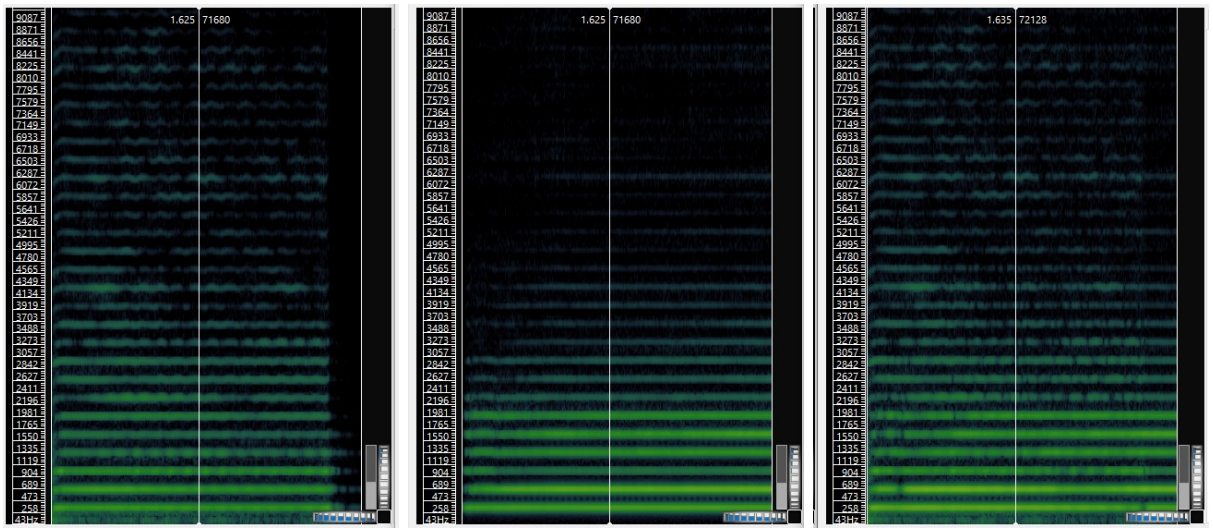
Treino	Neurônios	SDR			Loss
		Fonte 1	Fonte 2	Média das Fontes	
INS1	5	-0,7631	7,4790	**3,3579**	1,70e-4
	100	-0,8258	7,1714	3,1728	**1,59e-4**
	500	-7,2995	4,5694	-1,3651	3,50e-04
INS2	5	0,0803	7,3460	**3,7131**	**2,75e-4**
	100	-0,4432	6,5662	3,0615	2,90e-4
	500	0,41721	4,2846	2,3509	3,74e-04
INS3	2	-1,3514	6,8746	2,7616	2,38e-4
	12	-2,7860	2,1704	-0,3078	6,14e-4
	22	-2,2206	7,2180	2,4987	4,45e-4
	32	-3,8710	7,5713	1,8502	3,52e-4
	42	-0,9473	9,2373	**4,1450**	**2,15e-4**
	52	0,3413	7,8903	4,1158	2,78e-4
INS4	10	-19,0496	20,7002	0,8253	**1,62e-4**
	15	-25,8182	20,0541	-2,8821	1,83e-4
	20	-22,3495	21,2168	-0,5663	1,63e-4
	25	-19,1608	0,3136	-9,4236	1,74e-4
	30	-17,2706	21,0512	**1,8903**	1,63e-4
	35	-27,7375	20,9781	-3,3797	1,74e-4
	40	-25,7064	21,1496	-2,2784	1,77e-4
	45	-23,5651	21,2707	-1,1472	1,75e-4
	50	-22,1805	21,1175	-0,5315	1,93e-4
INS5	14	-9,0494	7,0572	**-0,9961**	**2,74e-4**
INS6	14	-1,1561	7,7231	**3,2835**	**1,54e-4**
INS7	14	-10,0694	21,2625	**5,5965**	**1,46e-4**
	40	-12,1725	21,1705	4,4990	1,50e-4
INS8	20	14,1277	27,2977	20,7127	1,24e-04
	60	13,9812	31,9767	22,979	4,79e-05
	80	14,3394	26,6594	20,4994	1,43e-04
	100	13,0615	29,7743	**21,4179**	**6,80e-05**
	200	15,3552	24,8798	20,1175	2,25e-04

[4]Para referência das frequências das das notas musicais, consulte a Tabela I.1, nos Anexos.

Para cada quantidade de neurônios da configuração, mostra-se o valor de SDR da separação de cada um dos instrumentos (fonte 1 e fonte 2) e da média destes. Também é apresentado o valor do Loss, calculado com a Equação 4.1. Em negrito, estão sinalizadas a média de SDR dos melhores treinamentos e o menor Loss para cada configuração dentre as possíveis quantidades de neurônios.

Dos valores apresentados na tabela, o que mais chama a atenção é a diferença entre os valores de SDR do treino INS8 e os demais treinos, com a média de suas fontes sendo aproximadamente quatro vezes maior que a segunda maior média das fontes, do treino INS7. A avaliação da separação de cada fonte também é maior que em todos os treinos, com um aumento expressivo principalmente na fonte 1. Ressalta-se aqui que este é o único treinamento que utiliza como critério de parada a melhoria do SDR.

Na separação instrumental são apresentadas a seguir apenas os espectrogramas dos sinais, uma vez que suas formas de onda não são tão intuitivas de compreender em relação à separação com sinais artificiais. Mas antes de analisar o resultado da separação de sinais instrumentais, é necessário destacar algumas características dos mesmos que são visíveis em seus espectrogramas. A Figura 4.23 mostra os sinais originais e a mistura composta por contrabaixo e flauta na nota E4.



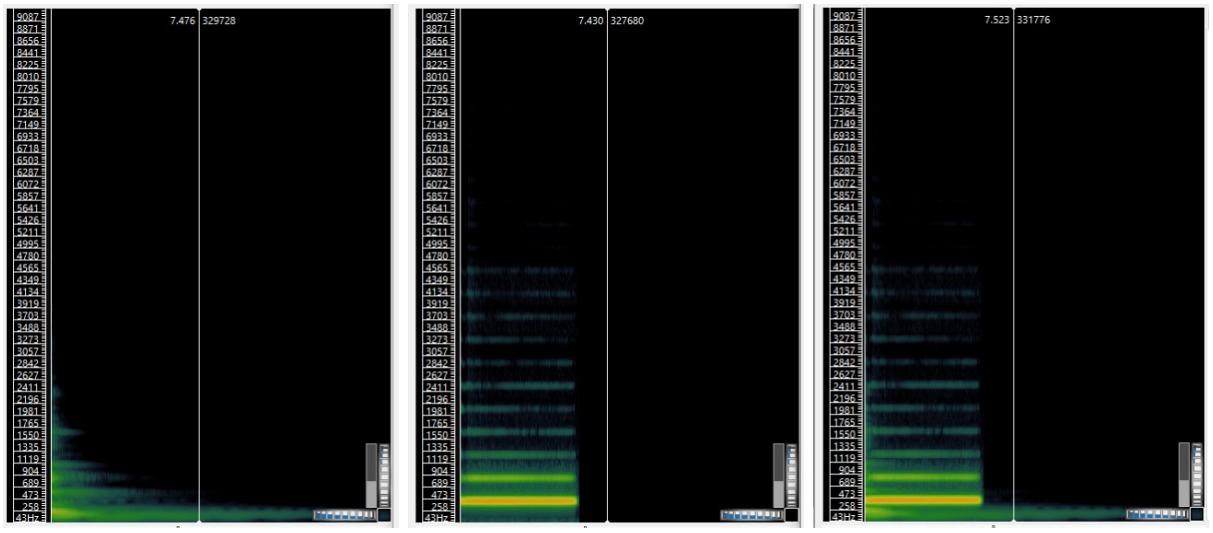
(a) Contrabaixo. (b) Flauta. (c) Mistura.

Figura 4.23: Espectrogramas dos sinais de contrabaixo e flauta puros e misturados.

Observando o espectrograma do contrabaixo (Figura 4.23a), vemos que este apresenta muitos harmônicos superiores, um ataque muito rápido e sustentação longa, no repouso seus harmônicos mais altos somem rapidamente e os de menor frequência silenciam-se pouco tempo depois, e durante todo o áudio há a presença de ruído de baixa frequência, boa parte proveniente do arco passando pelas cordas do instrumento (tal ruído é perceptível ao escutar o áudio). Já no espectrograma da flauta (Figura 4.23b), percebe-se menos harmônicos superiores em relação ao contrabaixo, não há o mesmo ruído de baixa frequência, seu ataque é rápido, sua sustentação é longa e, como o sinal fora cortado para manter o mesmo tamanho do contrabaixo, o repouso se dá

de forma instantânea. Devido ao fato de ambos sinais possuírem mesma frequência e sustentação longa, torna-se difícil distinguir a contribuição de cada uma no sinal misturado (Figura 4.23c), mas pode-se destacar que há o ruído de baixas frequências e os harmônicos de alta frequência do contrabaixo, e as altas frequências do sinal da flauta continuam atém o final do sinal, com repouso instantâneo.

Os espectrogramas dos sinais de harpa e trompa são apresentados na Figura 4.24.

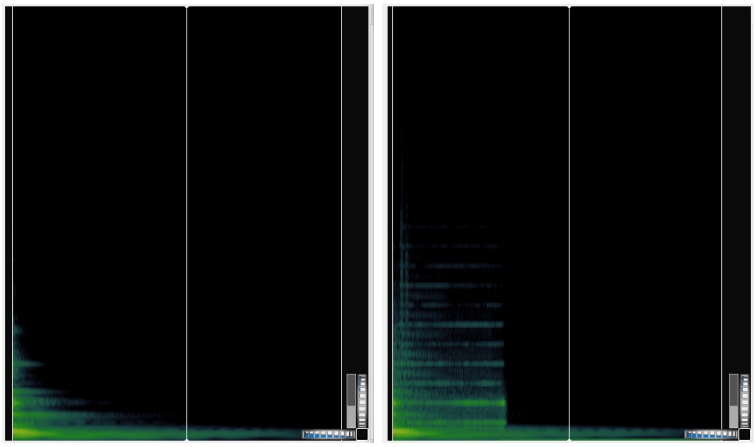


(a) Harpa. (b) Trompa. (c) Mistura.

Figura 4.24: Espectrogramas dos sinais de harpa e trompa puros e misturados.

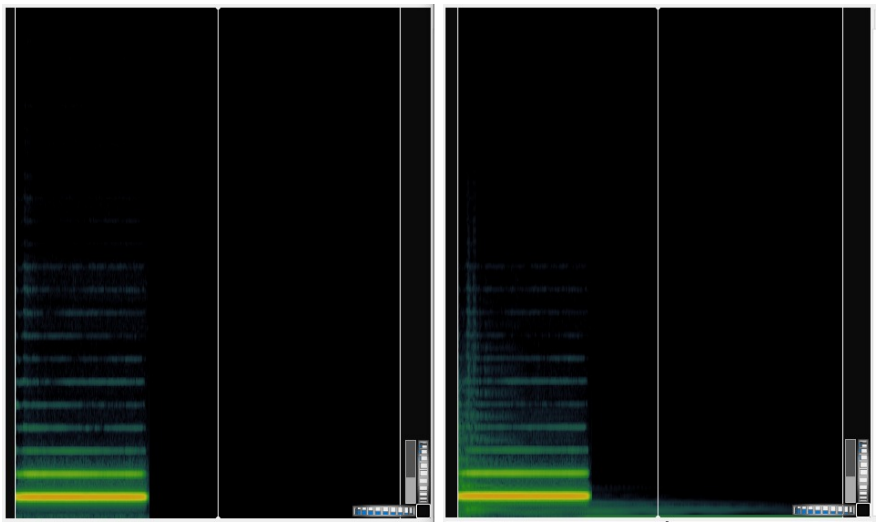
Estes sinais são mais facilmente diferenciáveis pelos seus espectrogramas, por possuírem uma envoltória e frequência fundamental consideravelmente diferentes. A harpa possui ataque muito rápido, não há sustentação e seu repouso demora bastante tempo com a reverberação da corda, enquanto a trompa também possui ataque muito rápido, mas um bom período de sustentação e seu repouso é muito rápido, além de ter presente mais harmônicos superiores em sua sustentação. O sinal da trompa fora estendido com silêncio para que tivesse o mesmo tamanho do sinal da harpa.

As Figuras 4.25 e 4.26 mostram os espectrogramas das saídas do treino INS8 com 100 neurônios.



(a) Esperado. (b) Obtido.

Figura 4.25: Espectrogramas dos sinais esperado e obtido de harpa do treino INS8 de 100 neurônios.

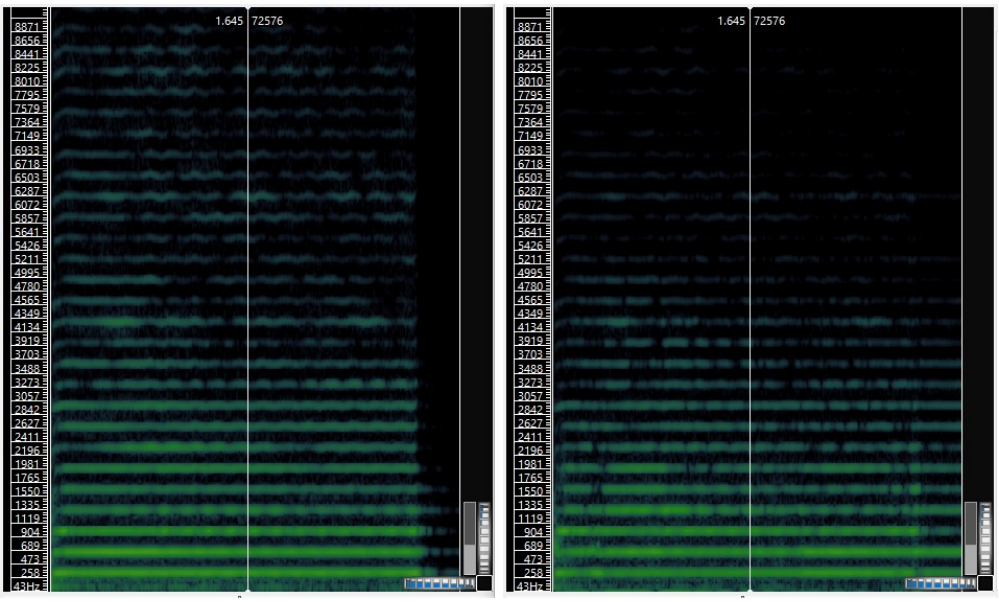


(a) Esperado. (b) Obtido.

Figura 4.26: Espectrogramas dos sinais esperado e obtido de trompa do treino INS8 de 100 neurônios.

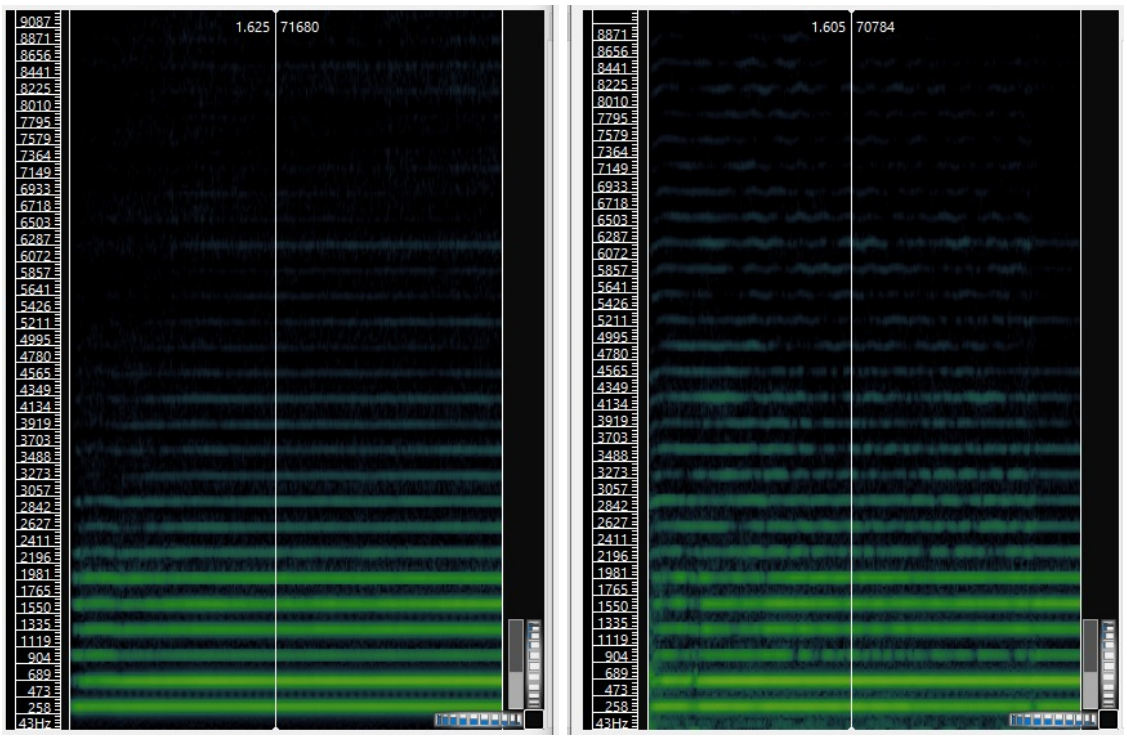
Em ambos espectrogramas, pode-se perceber que os sinais dos dois instrumentos estão presentes em cada saída. Entretanto, na saída destinada à harpa, o sinal da trompa está com amplitude bastante reduzida, e o contrário ocorre na saída destinada à trompa. O mesmo pode ser percebido ao escutar o sinal de áudio. Isto indica que a rede está no caminho correto para isolar uma fonte em relação a outra. Uma possível abordagem para melhorar a separação, considerando que esta está sendo feita de forma especializada, é exigir que mais épocas sejam treinadas. Um fato a se ressaltar é que o treino com 60 neurônios apresenta espectrogramas semelhantes aos que acabaram de ser apresentados, mas em seu áudio pode-se já perceber a presença de certas distorções, o que prejudica sua qualidade.

Dentre os resultados apresentados na Tabela 4.9, o que apresentou maior SDR para separação de contrabaixo e flauta foi o INS3 com 42 neurônios. Os espectrogramas de suas saídas podem ser vistos nas Figuras 4.27 e 4.28, junto com uma comparação da saída esperada.



(a) Esperado. (b) Obtido.

Figura 4.27: Espectrogramas dos sinais esperado e obtido de contrabaixo do treino INS3 de 42 neurônios.

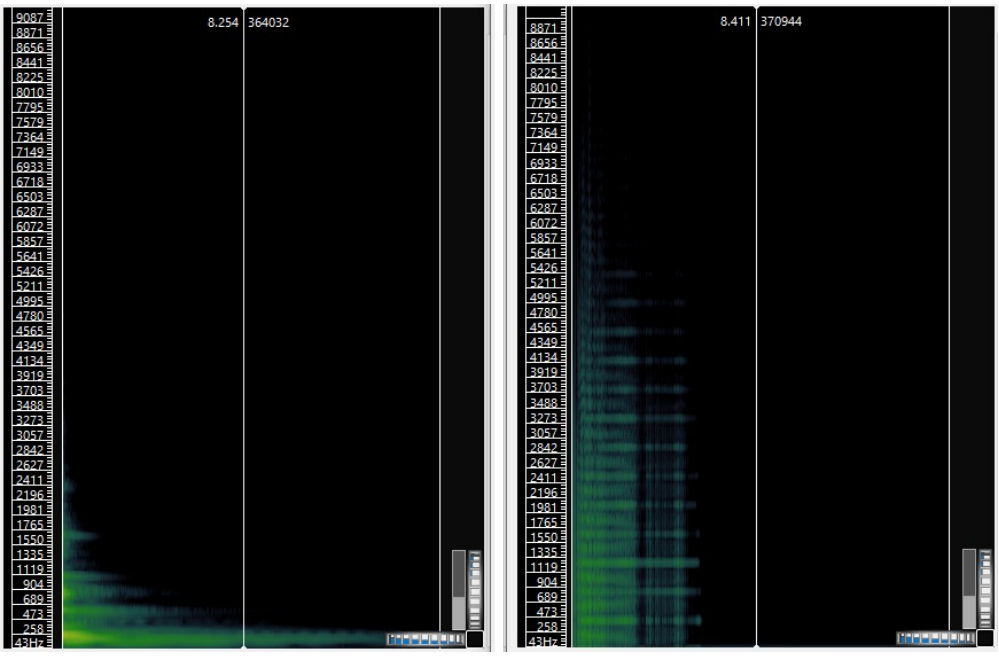


(a) Esperado. (b) Obtido.

Figura 4.28: Espectrogramas dos sinais esperado e obtido de flauta do treino INS3 de 42 neurônios.

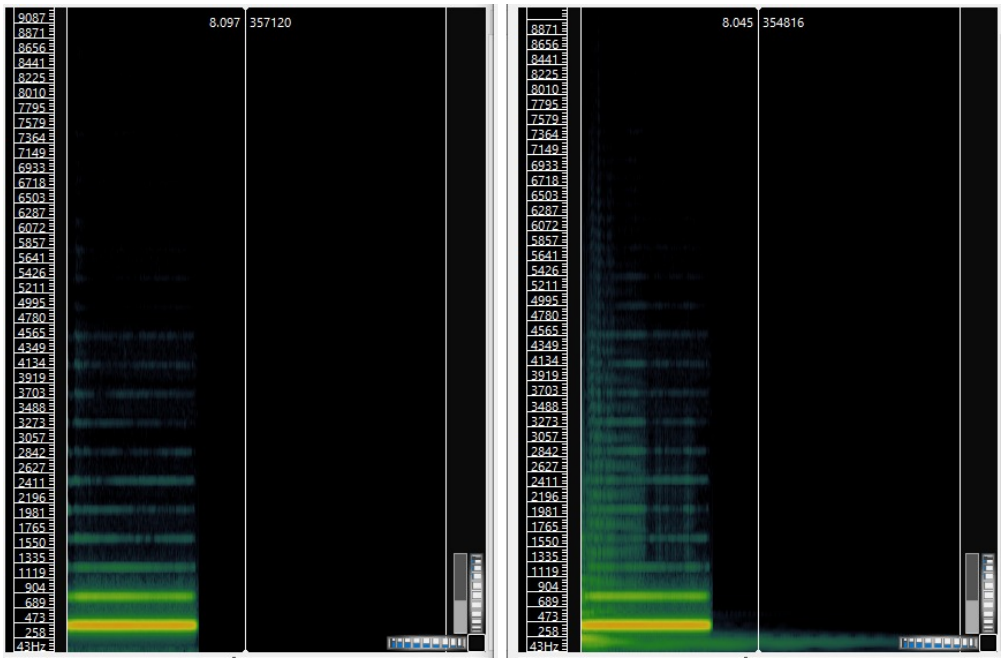
Devido à semelhança entre os espectrogramas de contrabaixo e flauta, torna-se difícil avaliar a separação na saída do canal do contrabaixo. Mas pode-se perceber que há menos harmônicos superiores que o esperado, o repouso é instantâneo (semelhante à flauta) e tem o ruído de baixa frequência ao longo do sinal, como na saída esperada. Já no canal da flauta, seu espectrograma possui ruído de baixa frequência igual ao esperado no contrabaixo, há mais harmônicos superiores que o sinal original e o repouso é instantâneo. Comparando as duas saídas, nota-se que as duas são muito semelhantes. Escutando os dois sinais de áudio, percebe-se que há a presença tanto do contrabaixo quanto da flauta em ambos. Porém, diferentemente da separação do treino INS8 com 100 neurônios, em cada canal a amplitude de ambas fontes é muito semelhante, dificilmente podendo-se afirmar que a separação fora feita com sucesso.

Para analisar os resultados da separação feita pela rede de Elman, são apresentados os resultados do treino INS7 de 14 neurônios, com os espectrogramas seus sinais de saída nas Figuras 4.29 e 4.30.



(a) Esperado. (b) Obtido.

Figura 4.29: Espectrogramas dos sinais esperado e obtido de harpa do treino INS7 de 14 neurônios.



(a) Esperado. (b) Obtido.

Figura 4.30: Espectrogramas dos sinais esperado e obtido de trompa do treino INS7 de 14 neurônios.

Nesta separação, além de ambas fontes estarem presentes nas duas saídas com amplitudes consideráveis, percebe-se ainda a adição de ruídos de diversas frequências. Ao escutar os áudios das saídas, nota-se que em ambos há a presença dos dois instrumentos, mas com uma distorção forte.

Estes foram os principais testes realizados nos sistemas das Figuras 3.1a e 3.2a. Em cada um foram extraídas informações importantes que podem orientar trabalhos futuros na área, apresentando métodos promissores em alguns casos, e, mesmo que algumas das propostas de separação sonora não alcançaram os objetivos propostos, ainda foi possível aprender com os mesmos, com as limitações encontradas neles. Existem inúmeros testes que ainda podem ser feitos, muitos caminhos a serem explorados, como novas arquitetura de redes neurais, sinais de áudio utilizados, métodos de segmentação, dentre outros. Os testes aqui feitos cumpriram seu propósito, trazendo clareza para alguns aspectos da área.

Capítulo 5

Conclusões

Neste trabalho foi feito um estudo a respeito da classificação e separação de fonte sonora online usando redes neurais, a partir de suas amostras no domínio do tempo. Redes recorrentes rasas de Elman e profundas LSTM foram avaliadas em testes com sinais artificiais e de instrumentos reais.

Para o caso do sistema de classificação, foi mostrado que o vetor de descritores proposto é suficiente para alcançar uma acurácia alta. Além disso, foi feita também uma comparação entre uma rede *feedforward* e uma recorrente, revelando que esta segunda possui mais robustez para classificação de sinais de maiores complexidade e que apresentam padrões temporais marcantes. Em comparação com os resultados publicados por Anderson [8], acurácia de 75%, observa-se que a rede de Elman com 40 neurônios na camada intermediária atingiu acurácia maior de 96,44% para o sistema proposto neste trabalho.

No sistema de separação, um ponto que fica claro é a complexidade do problema. Foi usado o SDR como figura de mérito para avaliar a qualidade da separação feita pelas redes. No caso mais simples, com ondas artificiais (onda senoidal, quadrada e triangular) e treinamento especializado, o sistema fornece uma saída boa, atingindo SDR de 51,05 . Porém, aumentando-se a complexidade visando buscar um treino generalizado, o desempenho cai fortemente para 4,9691 usando redes LSTM com 500 células de memória na camada escondida. Quando os sinais que se deseja separar são amostras de dois instrumentos musicais, percebe-se também que o desempenho da rede reduz para 21,4179 no melhor caso.

Dentre os métodos utilizados na separação, um que merece destaque é a utilização da melhoria do SDR como critério de parada. É fácil compreender que estas redes poderiam apresentar tal medida de desempenho maior que as demais, visto que é o buscado durante o treinamento, mas as análises feitas sobre as formas de onda, espectrogramas e ao escutar os áudios de saída mostram que de fato a separação dessas redes tem maior qualidade. É necessário notar também que este parâmetro não é trivial de se interpretar, considerando que nos experimentos realizados neste trabalho as melhores separações alcançaram valores de SDR na faixa de 40, ao passo que os sistemas Wave-U-Net e Demucs conseguiam valores em torno de 5, o que é considerado estado da arte na área. Deste modo interpreta-se que os valores de SDR dependem significativamente dos sinais sonoros em que é aplicado e suas complexidades. Outro ponto a considerar no uso do SDR

como critério de parada é que o treinamento visa reduzir o Loss (MSE) sem, no entanto, garantir que isso resulte em um aumento do valor de SDR.

5.1 Perspectivas Futuras

Em relação à classificação de fonte sonora, os resultados foram todos satisfatórios. Portanto, podem ser feitos testes mais complexos, aumentando a quantidade de notas na entrada ou até mesmo aplicando sinais em que o instrumento toca mais de uma nota ao longo do tempo, analisando a acurácia da rede nesses casos. Também espera-se desenvolver uma rede capaz de classificar mais do que dois instrumentos, explorando também instrumentos com sonoridade semelhante.

Para a separação de fonte sonora feita com sinais artificiais e de forma generalizada, uma possível abordagem para aprimorar a qualidade da separação da rede é com a aplicação de um conjunto de validação durante o treinamento da mesma. Espera-se também que sejam realizados testes das redes em bancos de dados mais comuns a esse problema.

Além disso, como os treinamentos parecem chegar nos limites de desempenho de uma rede com apenas uma camada LSTM, espera-se ainda que sejam feitos experimentos explorando um aumento na quantidade de camadas LSTM, além de acrescentar outras camadas, como a convolucional 1D (utilizada tanto na Wave-U-Net[9] quanto no Demucs[4]). A área de estudo é relativamente nova e existe uma gama enorme de abordagens que podem ser tomadas.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] REN, G. et al. A modified Elman neural network with a new learning rate scheme. *Neurocomputing*, v. 286, p. 11–18, 2018.

[2] KUMAR, S.; SUBHA, D. Prediction of depression from EEG signal using long short term memory(LSTM). In: *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. [S.l.: s.n.], 2019. p. 1248–1253.

[3] JEYALAKSHMI, C.; MURUGESHWARI, B.; KARTHICK, M. HMM and K-NN based automatic musical instrument recognition. In: *2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*. [S.l.: s.n.], 2018. p. 350–355.

[4] DÉFOSSEZ, A. et al. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019. Disponível em: <`https://arxiv.org/pdf/1911.13254.pdf`>.

[5] SANTINI, R. M.; SOUZA, R. F. d. Recuperação da informação de música e a ciência da informação: tendências e desafios de pesquisa. In: *VIII ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação*. [s.n.], 2007. Disponível em: <`http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/243`>.

[6] CHERRY, E. C. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 25, n. 5, p. 975–979, 1953.

[7] BAHRE, S.; MAHAJAN, S. P.; PILLAI, R. T. Novel audio feature set for monophonie musical instrument classification. In: *2017 International Conference on Recent Innovations in Signal Processing and Embedded Systems (RISE)*. [S.l.: s.n.], 2017. p. 562–565.

[8] ANDERSON, T.-A. Musical instrument classification utilizing a neural network. In: *2017 12th International Conference on Computer Science and Education (ICCSE)*. [S.l.: s.n.], 2017. p. 163–166.

[9] STOLLER, D.; EWERT, S.; DIXON, S. Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. In: *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*. [s.n.], 2018. Disponível em: <`http://arxiv.org/abs/1806.03185`>.

[10] SILVA, I. Nunes da et al. *Artificial Neural Networks - A Practical Course*. [S.l.]: Springer, 2017.

[11] MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.

[12] Wikipedia contributors. *Artificial neuron — Wikipedia, The Free Encyclopedia*. 2021. `https://en.wikipedia.org/w/index.php?title=Artificial_neuron&oldid=1012965643`. [Online; accessed 17-May-2021].

[13] ELMAN, J. L. Finding structure in time. *Cognitive science*, v. 14, n. 2, p. 179–211, 1990.

[14] SRINIVASAN, V.; ESWARAN, C.; SRIRAAM, N. Approximate entropy-based epileptic EEG detection using artificial neural networks. *IEEE Transactions on Information Technology in Biomedicine*, v. 11, n. 3, p. 288–295, 2007.

[15] AGGARWAL, C. C. *Neural Networks and Deep Learning - A Textbook*. [S.l.]: Springer, 2018.

[16] HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997.

[17] GREFF, K. et al. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, v. 28, n. 10, p. 2222–2232, 2017. ISSN 2162-2388. Disponível em: <`http://dx.doi.org/10.1109/TNNLS.2016.2582924`>.

[18] Wikipedia contributors. *Long short-term memory — Wikipedia, The Free Encyclopedia*. 2021. [Online; accessed 16-May-2021]. Disponível em: <`https://en.wikipedia.org/w/index.php?title=Long_short-term_memory&oldid=1022481054`>.

[19] Wikipedia contributors. *Hadamard product (matrices) — Wikipedia, The Free Encyclopedia*. 2021. [Online; accessed 18-May-2021]. Disponível em: <`https://en.wikipedia.org/w/index.php?title=Hadamard_product_(matrices)&oldid=1022283183`>.

[20] YADAV, S.; SHUKLA, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: *IEEE 6th International Conference on Advanced Computing (IACC)*. [S.l.: s.n.], 2016. p. 78–83.

[21] MIRANDA, E. *Computer Sound Design: Synthesis Techniques and Programming*. [S.l.]: Taylor & Francis, 2002.

[22] ROCCHESSO, D. *Introduction to Sound Processing*. Firenze, Italy: PHASAR Srl, 2003.

[23] Wikipedia contributors. *Spectrogram — Wikipedia, The Free Encyclopedia*. 2021. `https://en.wikipedia.org/w/index.php?title=Spectrogram&oldid=1001441730`. [Online; accessed 12-April-2021].

[24] SONIC Visualizer Homepage. `https://www.sonicvisualiser.org/`. Accessed: 2021-05-16.

[25] 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.

[26] SONIC Visualizer - Reference Manual. `https://www.sonicvisualiser.org/doc/reference/1.3/en/#waveform`. Accessed: 2021-05-16.

[27] MALT, M.; JOURDAN, E. Zsa. descriptors: a library for real-time descriptors analysis. In: *Proceedings of the 5th Sound and Music Computing Conference*. [s.n.], 2008. p. 134–137. Disponível em: <`https://hal.archives-ouvertes.fr/hal-01580326`>.

[28] FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

[29] VINCENT, E.; GRIBONVAL, R.; FEVOTTE, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 14, n. 4, p. 1462–1469, 2006.

[30] MANILOW, E.; SEETHARMAN, P.; SALAMON, J. *Open Source Tools & Data for Music Source Separation*. 2020. Disponível em: <`https://source-separation.github.io/tutorial`>.

[31] RAFII, Z. et al. *The MUSDB18 corpus for music separation*. dez. 2017. Disponível em: <`https://doi.org/10.5281/zenodo.1117372`>.

[32] LIUTKUS, A.; FITZGERALD, D.; RAFII, Z. Scalable audio separation with light kernel additive modelling. In: IEEE. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane, Australia, 2015. Disponível em: <`https://hal.inria.fr/hal-01114890`>.

[33] MATHWORKS - MATLAB. `https://www.mathworks.com/products/matlab.html`. Accessed: 2021-05-17.

[34] BAUCHSPIESS, D. *Repositório: Separação de Fonte Sonora*. Disponível em: <https://github.com/minibauchspiess/Separacao-de-Fonte-Sonora>.

[35] IRCAM Solo Instruments 2. `https://www.uvi.net/en/orchestral/ircam-solo-instruments-2.html`. Accessed: 2021-05-17.

[36] ZHANG, J. X.; WHALLEY, J.; BROOKS, S. A two phase method for general audio segmentation. In: *2009 IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2009. p. 626–629.

ANEXOS

	C	C#	D	Eb	E	F	F#	G	G#	A	Bb	B
0	16.35	17.32	18.35	19.45	20.60	21.83	23.12	24.50	25.96	27.50	29.14	30.87
1	32.70	34.65	36.71	38.89	41.20	43.65	46.25	49.00	51.91	55.00	58.27	61.74
2	65.41	69.30	73.42	77.78	82.41	87.31	92.50	98.00	103.8	110.0	116.5	123.5
3	130.8	138.6	146.8	155.6	164.8	174.6	185.0	196.0	207.7	220.0	233.1	246.9
4	261.6	277.2	293.7	311.1	329.6	349.2	370.0	392.0	415.3	440.0	466.2	493.9
5	523.3	554.4	587.3	622.3	659.3	698.5	740.0	784.0	830.6	880.0	932.3	987.8
6	1047	1109	1175	1245	1319	1397	1480	1568	1661	1760	1865	1976
7	2093	2217	2349	2489	2637	2794	2960	3136	3322	3520	3729	3951
8	4186	4435	4699	4978	5274	5588	5920	6272	6645	7040	7459	7902

Figura I.1: Notas musicais e suas frequências (em Hz), na escala temperada[1].

[1]Disponível em `https://www.researchgate.net/publication/323752411_Development_of_a_Novel_Method_for_Automatic_Detection_of_Musical_Chords/figures?lo=1`