UNIVERSIDADE FEDERAL DO RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

Estudo para Detecção de Eventos Sonoros como Comunicação de Alertas para Surdos

Douglas Severo Silveira

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Computação

Orientadora: Profa. Dra. Regina Barwaldt

Rio Grande, 2019

Ficha catalográfica

S587e Silveira, Douglas Severo.
Estudo para detecção de eventos sonoros como comunicação de alertas para surdos / Douglas Severo Silveira. – 2019.
93 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Computação, Rio Grande/RS, 2019.
Orientadora: Dra. Regina Barwaldt.

1. Detecção 2. Classificação 3. Eventos Sonoros 4. Alerta Visual 5. Surdez I. Barwaldt, Regina II. Título.

CDU 616.28-008.14:81’221

Catalogação na Fonte: Bibliotecário José Paulo dos Santos CRB 10/2344

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

DISSERTAÇÃO DE MESTRADO

Detecção de Eventos Sonoros como Comunicação de Alertas para Surdos

Douglas Severo Silveira

Banca examinadora:

Profª. Drª. Raquel de Miranda Barbosa

Prof. Dr. Vagner Santos da Rosa

Profª. Drª. Regina Barwaldt
Orientadora

RESUMO

SILVEIRA, Douglas Severo. **Estudo para Detecção de Eventos Sonoros como Comunicação de Alertas para Surdos**. 2019. 93 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

Com a evolução na área de tecnologia, soluções destinadas a pessoas com deficiência podem possibilitá-las viverem com maior independência, segurança e conectividade com o resto do mundo. Alguns eventos emitem sinais característicos que podem ser interpretados, por exemplo, como a ocorrência de uma situação de perigo. Os surdos, por não receberem os sinais sonoros, seguidamente mantém-se em estado de alerta, fazendo varreduras visuais nos ambientes, dificultando o processo natural de atenção seletiva e a concentração em outras atividades. Foi elaborado um levantamento prévio por meio de questionário estruturado, respondido por pessoas com surdez ou que possuem alguma relação, coletando informações sobre a demanda dos surdos, como a identificação de sinais de alerta importantes. Os resultados apontaram ser importante que o surdo tenha um recurso que auxilie na identificação de eventos que caracterizam situação de perigo. O desenvolvimento realizou estudos elucidativos sobre o público alvo/motivador, tecnologias existentes e comunicação visual. Desta forma, o trabalho tem como objetivo apresentar um estudo de caso de aplicação de modelo de rede neural profunda que, por meio de Detecção de Eventos Acústicos (AED), visa classificar alertas sonoros específicos para surdos. Foram realizados dois experimentos utilizando modelos de rede neural profunda, utilizando parte de três conjuntos de dados disponibilizados por grupo de pesquisa ligado a AED. Os dados foram sintetizados em um *dataset* que pode ser dividido entre classes de sons de alerta de sons ambiente. Um dos testes utilizou áudios brutos como entrada, os demais a extração do cepstrum de mel-frequência (MFCC), extraído em etapa de pré-processamento. Como resultados, quatro dos testes alcançaram *F1 Score* baseado em segmento acima de 85%, caracterizando possível problema de *overfitting* e outro alcançou 14% baseado em evento, caracterizando *underfitting*. Ao final, são discutidas possíveis causas para problemas apresentados, é sugerido o seguimento da pesquisa no modelo que utiliza uma Rede de Memória de Longo Prazo (LSTMs) e alterações no *dataset* para obtenção de melhores resultados.

Palavras-chave: Detecção, classificação, eventos sonoros, alerta visual, surdez.

ABSTRACT

SILVEIRA, Douglas Severo. **Detection of Acoustic Events as Communication of Alerts for Deaf People**. 2019. 93 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

The evolution in the area of technology possibility solutions for people with disabilities to improve our independence, security and connectivity with the rest of the world. Some events emit characteristic signals and it can be interpreted, for example, as the occurrence of a dangerous situation. The deaf, because they did not receive the sound signals, then kept on alert, making visual sweeps in the environments, making it difficult for the process of natural selective attention and concentration on other activities. A previous survey was conducted using a structured questionnaire, answered by people with deafness or who have some relation, collecting information about the deaf people's demand, such as the identification of important warning signs, the results pointed out importance to deaf people a resource to assists identification for events of danger situations. The development carried out elucidative studies on the target/motivating public, existing technologies and visual communication. This work presents the use case application of a deep neural network model in Acoustic Events Detection (AED) to classify specific sound alerts for deaf people. Two experiments were carried out to adapt deep neural network models, using part of three datasets provided by a research group linked to AED. The data has been synthesized in a dataset that can be divided between ambient sound alert sound classes. Each model sent information to its neural network in a format different from the same resources. One of the tests used raw audios as input, the others the extraction of honey-frequency cepstrum (MFCC), extracted in pre-processing stage. As a result, four of the tests achieved segment-based F1 Score above 85%, characterizing possible overfitting problem and another reached 14% event-based characterizing possible underfitting. In the end, possible causes for problems presented are discussed, it is suggested to follow the research in the model that uses a Long Term Memory Network (LSTMs) and changes in the dataset for better results.

Keywords: Detection, classification, acoustic events, visual alert, deafness.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AASP - *Audio e Acoustic Signal Processing*

AED - Detecção de Eventos Acústicos

ALD - Dispositivo Auxiliar de Audição

ANN - Rede Neural Artificial

BLSTM - *Bidirectional Long Short-Term Memory Networks*

CNN - Rede Neural Convolucional

CRNN - Redes Neural Convolucional Recorrente

DCT - Transformada Discreta de Cosseno

DL - *Deep Learning*

DNN - Redes Neurais Profundas

ES - Ensino Superior

FDP - Função Densidade de Probabilidade

FENEIS - Federação Nacional de Educação e Integração dos Surdos

FFT - Transformada Rápida de Fourier

FNN - *Feedforward Neural Network*

GMM - *Gaussian Mixture Model*

GD - Gradiente Descendente

HMM - *Hidden Markov Model*

HSL - *Hue Saturation Lightness*

HT - Transformada de Hough

IA - Inteligência Artificial

IEEE - *Institute of Electrical and Electronic Engineers*

LIBRAS - Língua Brasileira de Sinais

LSTM - Rede de Memória de Longo Prazo

MEC - Ministério da Educação

MFCC - *Mel Frequency Cepstral Coefficient*

ML - *Machine Learning*

MLP - *Multi Layer Perceptron*

MobNet - *Mobile Network*

PLN - Processamento de Linguagem Natural

RELU - Função Linear Retificada

RNN - Rede Neural Recorrente

SED - Detecção de Eventos Sonoros

SGD - Gradiente Descendente Estocástico

SVM - *Support Vector Machine*

W3C - *World Wide Web Consortium*

WDF - Federação Mundial de Surdos

SUMÁRIO

1 INTRODUÇÃO

Antigamente as pessoas com perda auditiva, por vezes, dependiam dos outros para a sua segurança e conectividade com o resto do mundo. Atualmente, com a evolução da tecnologia, também existem dispositivos que ajudam as pessoas com perda auditiva a viverem com maior independência (JONES, 2017). No âmbito brasileiro, segundo dados do Censo de 2010 realizado pelo IBGE, 9,7 milhões de pessoas têm deficiência auditiva, ou seja, 5,2% da população. Desses, 2.147.366 foram declarados com deficiência auditiva severa, situação em que há uma perda entre 70 e 90 decibéis (dB) (IBGE, 2010). Percebe-se, então, que grande parcela da população brasileira está em situação de dificuldade para ter uma vida independente e nestes casos a tecnologia pode oferecer, por meio da promoção ou potencialização de funções similares às do corpo humano, alternativas para a melhoria da vida dessas pessoas.

Por possuirmos mecanismos genéricos de seleção, um estímulo, por exemplo auditivo, pode nos alertar, fazendo despertar ou provocar respostas instintivas. Já a falta de atenção em sala de aula pode causar a falha de metodologias de ensino como a de "mostrar e dizer". Manter a sala isolada e livre de distrações, que não haja nada mais para ser visto ou ouvido, é uma estratégia para que o estudante preste atenção, o silêncio é quase sempre a regra (SKINNER, 1972).

Visto que os surdos são principalmente seres visuais, sendo seus olhos o portal para o mundo da informação e do conhecimento (WDF, 2007), há alguma tecnologia que possa ser utilizadas para auxiliar os mecanismos de seleção das pessoas surdas no seu dia a dia? No ambiente educacional há uma escassez de materiais didáticos e, também, dificuldades das escolas em tentar de torná-los mais agradáveis. Mesmo que esforços na criação de materiais que interessem mais - como interiores coloridos, mobília confortável e arranjos que conduzam à sociabilidade - não contribuam diretamente para o ensino daquilo que os estudantes devem aprender na escola, são reforçadores que contribuem para o ambiente da instrução, fortalecendo uma atitude positiva em relação à escola (SKINNER, 1972).

No âmbito das pessoas surdas, a inexistência de um sentido que possibilita a captação e, por consequência, interpretação de signos sonoros afeta a forma de configurar a realidade em relação a pessoas ouvintes. Por conta disso, necessitam utilizar de outros

sentidos, como a visão, para manterem-se alertas aos acontecimentos de qualquer ambiente. Imaginar que um deficiente auditivo não corre riscos porque a deficiência auditiva só o afeta no que se refere a comunicação interpessoal, é ter uma visão limitada dessa realidade. Mesmo que a sobrevivência dele seja menos ameaçada, não significa que um deficiente auditivo seja totalmente independente, diversos riscos à integridade física podem ocorrer, por exemplo, quando o surdo não percebe um alarme de incêndio, ou um carro ao atravessar a rua. (AUSTREGESILO, 2014).

No artigo de ZOVICO (2012) o autor apresenta aspectos da evolução das tecnologias de apoio a surdos e um dos apontamentos é o fator histórico analisado naquela pesquisa desde a década de 90, que torna inegável a dificuldade que essas pessoas vêm passando em suas vidas. Por conta disso, aponta o autor, que são necessários mais avanços com intuito de atender e incluir mais essas pessoas na sociedade. A possibilidade de estudo de uma tecnologia que possa vir a melhorar o dia a dia de pessoas surdas é o que mais motiva o objeto desta dissertação de mestrado.

A ideia de uma tecnologia que ofereça benefícios a pessoas com problemas de audição surgiu a partir de relato de um professor, surdo congênito, então primeiro vice-presidente da diretoria corporativa da Federação Nacional de Educação e Integração dos Surdos (Feneis)[1], que ministrava aula de Língua Brasileira de Sinais (LIBRAS) ao pesquisador em curso de formação de professores. Atuante junto à comunidade surda e, de acordo com a definição da Feneis, preocupado em estimular a autonomia pessoal, a interação e o contato com expressões e modos diversos de pensar, agir e sentir, este professor relatou em sala de aula, diversas dificuldades enfrentadas pela necessidade de estar sempre atento visualmente a qualquer ambiente. Apontou que há grande dificuldade em perceber o que ocorre fora do campo de visão central ou enquanto está exercendo outras atividades que lhe exijam atenção, como realizar atividades no computador.

Neste contexto, uma tecnologia capaz de identificar sons no ambiente e emitir alertas pode auxiliar o mecanismo de seleção que atua sob a atenção das pessoas surdas podendo, assim, beneficiar o processo de aprendizagem dessas pessoas. A Detecção de Eventos Acústicos (AED) tem por objetivo a análise sonora para detecção de eventos em sons polifônicos. Abordagens mais tradicionais nesta área utilizam métodos como Mel Frequency Ceps-tral Coefficient (MFCC), Gaussian Mixture Model (GMM's), Hidden Markov Models (HMM's). Outras formas utilizam a análise de espetogramas, como transformações de Hough (HT) e, mais recentemente e com melhores resultados, Redes Neurais Artificiais (ANN) do tipo Feedforward Neural Networks (FNN's).

As Redes Neurais Artificiais (ANN) são modelos computacionais apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que, por meio

[1]Feneis é uma entidade filantrópica, sem fins lucrativos, que tem por finalidade a defesa de políticas linguísticas, educação, cultura, saúde e assistência social, em favor da comunidade surda brasileira, bem como a defesa de seus direitos. Mais informações ⟨http://feneis.org.br/sobre/⟩.

de neurônios com pesos e possibilidade de configurações, tem a possibilidade de adquirir conhecimento através da experiência/aprendizado. Este trabalho realizou o treinamento de diferentes algoritmos de Redes Neurais Artificiais, seguido de testes simulados nestas redes e analisou os resultados obtidos comparando taxa de acerto da redes para cada tipo de som analisado, proximidade da arquitetura com uma situação real de utilização e apontou aspectos que podem ter influenciado positiva e negativamente nos resultados.

Demonstra aderência do tema ao grupo de estudos em Tecnologias Educacionais da linha de pesquisa Informática na Educação - FURG[2], contribuindo cientificamente, principalmente, para a comunidade do Centro de Ciências Educacionais (C3)[3] da Universidade Federal do Rio Grande (FURG)[4], visto que não são conhecidos estudos neste centro que abordem, como este, problemas de Detecção de Eventos Acústicos (AED). Sua relação com o ensino e a intenção de desenvolvimento de uma solução para o dia a dia de pessoas surdas apontam relevância social, que é justificada em relatos obtidos por pessoas que têm alguma relação com surdez e em estudos que apontaram potencial para auxiliar os estudantes em momentos de aprendizagem. Além disso, o projeto tem caráter inovador quando na aplicação de tecnologias de AED para área da surdez dá subsídio inicial, incitando o surgimento de uma solução relatada de importância aos surdos, diferente das encontradas em pesquisa bibliográfica e que o público alvo também não tem conhecimento da existência de algo parecido.

Como ponto inicial do desenvolvimento, apoiada por relatos e leituras anteriores, foi elaborada uma proposta de aplicação para detecção de eventos sonoros e apresentada à pessoas que possuem algum tipo de relação com a surdez por meio de instrumento de coleta (Apêndice A). Desta forma, foram obtidos dados que apontaram ser importante o desenvolvimento de um sistema de alerta para surdos que identifique determinados padrões de sons e informe, em linguagem visual, a ocorrência de eventos.

Por meio de abordagem baseada no resultado da aplicação de rede neural, este estudo apresenta uma proposta de sistema de alerta para surdos que identifica determinados padrões de sons, podendo, desta forma, ser utilizado para contribuir no processo de seleção de informações da pessoa surda, auxiliando no aproveitamento de sua atenção.

1.1 Objetivo Geral

O presente trabalho tem como objetivo geral apresentar um estudo de caso para detecção, classificação e comunicação para sugerir a ocorrência de eventos sonoros a pessoas surdas.

[2]Grupo de pesquisa Tecnologias Educacionais do Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande (FURG). Mais informações em ⟨http://infoeduc.c3.furg.br/⟩.

[3]Centro de Ciências Educacionais da Universidade Federal do Rio Grande (FURG). Mais informações em ⟨http://c3.furg.br/⟩.

[4]Universidade Federal do Rio Grande (FURG). Mais informações em ⟨http://furg.br/⟩.

1.1.1 Objetivos Específicos

- Elucidar sobre a pessoa surda e sua relação com a educação;
- Estudar formas de alerta visual;
- Identificar tipos de eventos sonoros;
- Estudar tecnologias para a detecção e classificação de eventos sonoros;
- Prototipar uma solução de detecção de tipos de sons estabelecidos;
- Analisar e documentar os resultados obtidos.

2 EMBASAMENTO TEÓRICO

A fundamentação deste estudo foi dividida em três partes, observando três aspectos: conhecimento do público alvo, motivador; tecnologias existentes que viabilizem implementação; e comunicação de forma visual. Esses aspectos correspondem a três seções, apresentadas abaixo:

- Público alvo, motivador: Surdez e Educação (seção 2.1), busca elucidar sobre a pessoa e cultura surda e sua relação com o ensino.
- Tecnologias existentes: Detecção de Eventos Sonoros (seção 2.3), apresenta arquiteturas que utilizam redes neurais artificiais aplicadas a problemas de Detecção de Eventos Acústicos (*Acoustic Event Detection* - AED).
- Comunicação visual: Linguagem Visual (seção 2.2), elucida sobre estratégias gráficas para a chamada de atenção e comunicação visual do sistema com o usuário.

A pesquisa objetivou contribuir para a melhor compreensão do tema estudado, apontar as iniciativas que estão sendo desenvolvidas, avanços da área, grupos e autores representativos (PIZZANI L.; ROSEMARY, 2012). Nas seções que seguem serão apresentados os resultados de buscas na literatura relacionadas à pessoa surda e educação, detecção de eventos sonoros e linguagem visual que desenvolvem em perspectivas e problematizações correlatas à exposta neste trabalho.

Este trabalho utilizou três estratégias de busca com termos em inglês contendo operadores lógicos e símbolos de truncagem e, de forma a complementar a pesquisa, estratégias em português correspondentes. Para exemplificar o entendimento objetivado na estratégia, as versões em inglês estão listadas abaixo juntamente com expectativa que os resultados possibilitem:

- `(deafness OR deaf) AND education`: compreender a surdez, aspectos de pessoas surdas em ambiente educacional
- `(sound OR acoustic) AND event* AND (detection OR classification) AND (deep learning)`: compreender a detecção

de eventos sonoros e identificar formas de implementação utilizando redes neurais profundas

- `visual AND language`: identificar formas de apresentação de alertas visuais

A síntese das leituras do produto destas buscas são apresentados nas seções seguintes deste capítulo.

2.1 Surdez e Educação

Como o projeto que intenciona beneficiar prioritariamente pessoas surdas no ambiente educacional, este capítulo aborda alguns aspectos peculiares da cultura surda, com propósito de conhecer melhor suas relações sociais, ambientais e questões educacionais que contribuam para o desenvolvimento de uma proposta de tecnologia que ajude o surdo no processo de aprendizagem.

2.1.1 Surdez

A surdez se caracteriza por uma dificuldade na recepção, percepção e reconhecimento de sons. Podendo ser chamada, também, de deficiência auditiva, pode ocorrer em diferentes graus, do mais leve (que não impede a comunicação oral) ao mais profundo (LIMA et al., 1997). A classificação da perda auditiva, representada na Tabela 1, é baseada na média dos limiares das frequências de 500, 1000 e 2000 Hz, que são consideradas as frequências das fala (SILMAN, 1991).

Tabela 1: Graus de perda auditiva

Classificação	Média dos limiares da fala
Normal	até 25 dB
Leve	de 26 a 40 dB
Moderada	de 41 a 55 dB
Moderadamente severa	de 56 a 70 dB
Severa	de 71 a 90 dB
Profunda	maior que 91 dB

Fonte: SILMAN (1991)

Nesta classificação, surdo é aquele que tem perda auditiva profunda e que, portanto, dificilmente adquirirá linguagem oral sem um treinamento específico para utilização da audição residual da fala. A pessoa que possui uma perda auditiva leve ou moderada ainda é capaz, por exemplo, de ouvir uma pessoa falando, já para uma perda moderadamente severa isto não é possível, mas ainda pode ouvir um bebê chorando, aspirador ou cachorro latindo, sons estes que não são audíveis por quem tem perda severa (MOURA, 2016).

Apesar de não ser possível identificar a surdez antes do nascimento, ela pode ocorrer ainda no período pré-natal, durante a gravidez, pelo uso de entorpecentes, alcoolismo

e exposição radiação por parte da mãe, por doenças que acometem a mãe, como, por exemplo, rubéola, meningite, herpes, entre outras, ou mesmo por fatores hereditários. O fator hereditário não é absoluto, na mesma família, por exemplo, pode haver surdos e ouvintes. Relacionados ao parto e ambiente hospitalar estão outros fatores, como o parto prematuro, nascimento tardio do bebê, infecções hospitalares e falta de oxigênio no cérebro do bebê (MOURA, 2016).

Após o nascimento, no segundo ou terceiro dia de vida, o diagnóstico pode ocorrer por meio do teste da orelhinha[1], apontando suspeitas, que devem ser confirmadas, ou não, por volta de quatro meses de idade com o apoio de outros testes. Porém, em muitos casos, a surdez só é percebida e diagnosticada por volta dos cinco anos de idade. Nesse período, é recorrente, a família iniciar uma corrida a fim de compreender, aceitar e procurar alternativas educacionais e de saúde adequadas (MOURA, 2016).

2.1.2 A Pessoa Surda

Sem negar a falta de audição do corpo surdo, foi adotada a proposta de LOPES (2007) que busca olhar a surdez por outro prisma que não da deficiência, mas o da diferença cultural. Com isso, nesta seção são apresentadas características de elementos peculiares a pessoas surdas, como identidade e cultura surda, formas de comunicação e linguagem.

Analisar a surdez pela diferença cultural se faz necessária para esta pesquisa, pois a abordagem pela deficiência foi construída culturalmente por narrativas de campos discursivos distintos, tais como clínicos, linguísticos, religiosos, educacionais, jurídicos, filosóficos etc. que descrevem, muitas vezes, o sujeito com surdez e podem não ser suficiente para o entendimento do que é a pessoa surda em sua essência.

Inicialmente, é importante salientar que o termo "surdo", forma que a comunidade defende que seja utilizado, denomina indivíduos que fazem parte de uma comunidade, com cultura e língua própria. Este termo faz oposição aos termos "surdo-mudo" que pressupõe a inabilidade deles para a fala, e "deficiente auditivo", em que o *déficit* é predominantemente marcado. A utilização deste termo representa estarem sendo respeitados pontos básicos que interferem na formação de um indivíduo, que o caracterizam e o distinguem dentro da sociedade (TERRA, 2011).

O reconhecimento dessas pessoas como sujeitos surdos significa atribuir a eles características de uma comunidade específica que, sem oposição aos ouvintes, tem a surdez como um *marcador cultural primordial* em prol de causas de lutas comuns, que no decorrer da história consolidaram a formação de uma comunidade surda (LOPES, 2007). Essa comunidade, que tem como centro o conceito semântico da palavra (unidade comum), proporcionou à surdez ser pensada a partir de bases culturais e históricas e, como

[1] O Teste da Orelhinha, ou "exame de emissões otoacústicas evocadas", é um método para constatar problemas auditivos nos recém-nascidos, tem duração de 5 a 10 minutos consiste na produção de um estímulo sonoro e na captação do seu retorno por meio de uma sonda introduzida na orelhinha do nenê.

Figura 1: Comunidade Surda

Fonte: autor

apresentado na Figura 1, é recorrentemente qualificada por quem a pertence meio de diversos significados sobre sentimentos de pertencimento, partilha, comunhão, sociedade, identidade e segurança. (LOPES, 2007).

Figura 2: Língua Brasileira de Sinais - LIBRAS

Fonte: Faculdade Plus - ⟨www.faculdadeplus.edu.br/course/pos-graduacao-em-docencia-de-libras-com-enfase-na-traducao-e-interpretacao/⟩.

Entre as discussões da comunidade estão o sistema de significações para representar coisas e negociar sentidos sobre elas. Essas associações formam uma linguagem que pode contribuir ao aprendizado, ser utilizada na socialização e interação das pessoas com o mundo. O Brasil possui a Língua Brasileira de Sinais (LIBRAS), que é uma língua de sinais (gestual). A LIBRAS possui a representação conforme Figura 2, é usada pela maioria dos surdos dos centros urbanos brasileiros e legalmente reconhecida como meio de

comunicação e expressão (CARVALHO, 2007). É um sistema linguístico de transmissão de ideias e fatos, oriundos de comunidades de pessoas surdas do Brasil, assim como ocorre em outras linguagens, ocupou uma posição única no aprendizado humano, funciona como meio de armazenar e transmitir informações, é o veículo para o intercâmbio de idéias e meio para que a mente humana seja capaz de conceituar (DONDIS, 1991).

Como forma de expressar, a linguagem faz uso de diversos signos. A definição clássica de signos define como algo que é usado, referido ou tomado no lugar de outra coisa (*aliquid pro aliquo*), podendo ser como entidades onde sons, sequências de sons ou correspondências gráficas que estão ligadas com significados ou conteúdos. Dessa forma, sons emitidos em ambientes do dia a dia são instrumentos de comunicação e representação capazes de auxiliar na configuração linguística da realidade e distinção de objetos e eventos entre si.

Quando um signo sonoro, como um ruído muito alto ou inusitado, ocorre com o significado de situação de perigo, nossos mecanismos genéricos de seleção podem fazer despertar instintivamente a atenção para possíveis pontos de emissão (SKINNER, 1972; VILELA M.; KOCH, 2001).

2.1.3 Inclusão no Ensino

Segundo a Política de Direitos Educacionais para Crianças Surdas (WDF, 2007), os surdos são principalmente seres visuais, sendo seus olhos o portal para o mundo da informação e do conhecimento. Na relação com o meio atua a linguagem, que com base nos sentidos que damos às coisas, construímos nossas experiências cotidianas e interpretações sobre nós e os outros (LOPES, 2007). Para melhorar essas relações, a política da Federação Mundial de Surdos (*World Federation of the Deaf*, WFD) afirma que a linguagem de sinais e as estratégias visuais devem ser disponibilizadas aos surdos como um direito de nascença.

No Brasil, como uma estratégia para a comunicação com outras pessoas, o surdo pode utilizar a Língua Brasileira de Sinais (LIBRAS), oficializada como segunda língua no Brasil conforme lei federal 10.436, de 24 de abril de 2002 (BRASIL, 2002), regulamentada pelo decreto nº 5.626, de 22 de dezembro de 2005 (BRASIL, 2005), oficializada no Brasil. O decreto, em acordo com a política da WFD, afirma que o surdo "por ter uma perda auditiva, compreende e interage com o mundo por meio de experiências visuais".

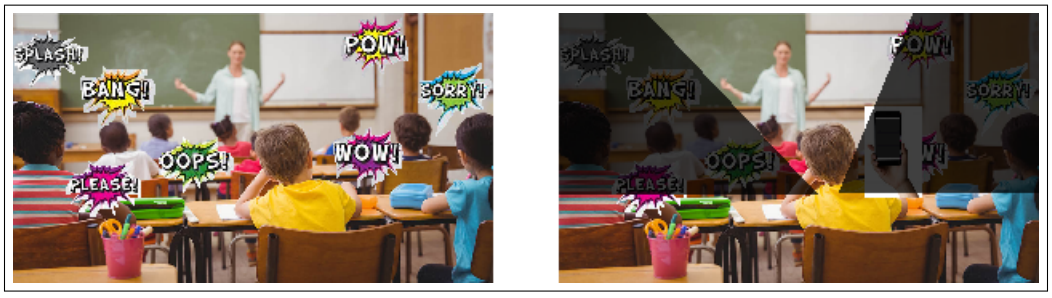
Além da oficialização da língua, o decreto nº 5.626 garante acesso dos estudantes surdos à escola regular, a inclusão da LIBRAS como disciplina curricular e outras estratégias, como formação de profissionais, que visam melhor estruturação desses ambientes. De forma complementar, para o ingresso ao ensino técnico e superior foi sancionada em de 28 de dezembro de 2016 a lei federal 13.409 (BRASIL, 2016) que inclui pessoas com deficiência no sistema de cotas. No ano posterior a esta lei, segundo o Ministério da Educação (MEC) (MEC/INEP/DEEP, 2018), houveram 2.138 matrículas de estudantes

surdos nos cursos de graduação em 2017.

Para DAROQUE S. C.; PADILHA (2011), a inclusão destes estudantes muitas vezes pode gerar conflitos, desafios, e ansiedade, tanto para a instituição e seus educadores, quanto para os próprios surdos que, apesar de inúmeras propostas e movimentações a inclusão no Ensino Superior (ES), principalmente ao acesso, ainda é um grande desafio para que as necessidades específicas desses estudantes sejam atendidas e para que se possa dizer que a inclusão é realidade.

Uma das estratégias que podem ser utilizadas com intenção de melhorar o processo de aprendizagem no ambiente educacional é direcionar o foco da atenção dos estudantes às partes mais importantes das atividades. Isso pode ser alcançado mantendo a sala de aula isolada e livre de distrações (SKINNER, 1972; LADEWIG, 2000). Dependendo do tipo de habilidade que está sendo aprendida e o estágio de aprendizagem que o indivíduo se encontra, a carga nos processos da atenção pode aumentar ou diminuir (LADEWIG, 2000).

Figura 3: Distrações em sala de aula



Fonte: autor

Mesmo que a eficiência no processo de seleção e descarte por parte do estudante dependa de fatores individuais, o uso correto de estratégias de atenção seletiva facilita a seleção de informações relevantes e auxilia no descarte de informações irrelevantes às atividades (LADEWIG, 2000). Além disso, há questões irrelevantes ao processo educacional que podem compartilhar da atenção e dificultar a "performance" dos indivíduos, como representado na Figura 3, onde a figura da direita demonstra a utilização de uma estratégia tecnológica filtrando os sons ambiente podendo deixar o estudante mais à vontade em permanecer focado na fala do professor, por exemplo. No caso dos surdos, segundo o relato que motivou este estudo, além de um possível evento ser uma distração, a necessidade de periodicamente fazer uma varredura visual no ambiente com a finalidade de verificar se não está ocorrendo alguma situação adversa também ocupa certa quantidade de atenção do estudante.

Neste caso, o uso de tecnologias pode ser uma importante estratégia cognitiva quando

auxilia as pessoas a descartarem mais facilmente causas de distrações do meio ambiente. Esse tipo de tecnologia, enquanto eficiente em possibilitar maior enfoque de atenção dos estudantes aos aspectos relevantes ao aprendizado, deve equipar o ambiente de ensino. Segundo SKINNER (1972) não há razão para que a sala de aula seja menos equipada do que ambientes como uma cozinha, para a qual são produzidas tecnologias para geladeiras, fornos e liquidificadores. O autor não acredita que um país que produz milhões desses eletrodomésticos não tenha condições de dispor do equipamento necessário para educar seus cidadãos em alto nível de competência da maneira mais eficiente.

2.2 Linguagem Visual

A linguagem visual é o único meio de comunicação humana que não dispõe de um conjunto de normas e preceitos, metodologia ou sistema único com critérios definidos, tanto para a expressão quanto para o entendimento dos métodos visuais (DONDIS, 1991). Esta seção não pretende apresentar soluções simples ou absolutas que determinem a melhor forma de comunicação visual para um sistema de alertas, mas, explanar sobre uma variedade de métodos de composição e *design* que levem em conta a diversidade da estrutura do modo visual para que embasem a formulação de algumas diretrizes de criação de símbolos que possam contribuir para a tomada da atenção do usuário e a comunicação visual.

2.2.1 Percepção

Segundo ALEXANDRE D. S.; TAVARES (2007) a visualização é o primeiro componente do sistema sensorial, o sentido adquirido mais rapidamente pelo cérebro e que possui capacidade de ter como alvo, para além do que está focado, toda a cena visual que se encontra dentro do campo de visão. Ele começa a operar quando recebe um estímulo luminoso, captado pela retina e convertido em sinais elétricos que são conduzidos a uma área de processamento primário do cérebro para gerar as informações iniciais de cor, forma, distância, tonalidade e outras.

Uma das teorias da percepção mais largamente conhecida e adotada em várias áreas é a de Gestalt[2]. Suas leis estabelecem características que determinam a pregnância de uma imagem, a forma como os elementos constitutivos de uma imagem podem vir a ser percebidos em termos organizacionais (ALEXANDRE D. S.; TAVARES, 2007).

Como um dos principais temas trazido por Gestalt, o fator básico da pregnância - palavra alemã que significa “boa figura” ou “vigor” - está relacionado a atividade perceptiva quando os elementos presentes em determinado ambiente são projetados de maneira sufi-

[2] A Psicologia da Gestalt, também conhecida como gestaltismo, psicologia da boa forma, leis de gestalt, é uma corrente de pensamento dentro da psicologia moderna surgida na Alemanha nos princípios do século XX, introduzido pela primeira vez por Christian Von Ehrenfels. Entende-se geralmente como um processo de dar forma, de configurar o que é colocado diante dos olhos, exposto ao olhar.

cientemente forte para destacar-se na cena exterior daquilo que ocorre na cena interior e, por consequência, vistos da forma mais simples possível, havendo a rápida assimilação do ambiente ou do elemento (ALEXANDRE D. S.; TAVARES, 2007), como representadas na Figura 4 a imagem à esquerda versus a da direita, possui baixa pregnância.

Figura 4: Contraste em exemplos de Pregnância

Fonte: Wedding Brasil - ⟨http://blogweddingbrasil.com.br/a-gestalt-na-sua-fotografia/⟩.

Conforme apontado por ALEXANDRE D. S.; TAVARES (2007), para Gestalt uma imagem alcança um fator alto de pregnância seguindo os princípios de proximidade, semelhança, fechamento, simplicidade, continuidade e figura/fundo. A lei de proximidade estabelece que mesmo não possuindo grande similaridade entre si, os elementos que se encontram próximos tendem a ser agrupados perceptivamente num conjunto. Para lei da semelhança normalmente não se sobrepõe à proximidade, de acordo com ela os elementos que possuem características semelhantes, principalmente em termos de cor, forma e textura, tendem a ser agrupados em conjuntos. A lei de fechamento diz que elementos dispostos de maneira a formar um contorno fechado ou formas incompletas tendem a ganhar maior grau de regularidade ou estabilidade, podendo vir a ganhar unidade (tendência da percepção humana em perceber formas completas). Já na lei de simplicidade os elementos são percebidos mais facilmente quando apresentam simetria, regularidade e não possuem texturas. Quanto a continuidade aponta que a percepção humana tende a orientar os elementos que parecem construir um padrão ou um fluxo na mesma direção. E na lei de figura/fundo é posto que qualquer campo perceptivo pode dividir-se numa figura sobre um fundo por meio de características da figura, como tamanho, forma, cor e posição (ALEXANDRE D. S.; TAVARES, 2007).

WARE (2004) afirma que o sistema de visão humano é usualmente dividido em três fases: processamento paralelo, percepção de padrões e processamento sequencial dirigido, e em cada uma delas pode-se aplicar os princípios da percepção visual. O proces-

samento paralelo ocorre por bilhões de neurônios que trabalham em paralelo para extrair propriedades de baixo nível da cena visual em causa, como orientação dos contornos, cor, textura e padrões de movimento, determinando ao que se deve dar atenção. A segunda etapa, de percepção de padrões, é flexível e influenciada pelas informações disponibilizadas pela primeira etapa. De processamento mais lento, processos ativos decompõem o campo visual em regiões e padrões simples envolvendo a memória a longo prazo, como contornos, regiões de cor semelhante e de textura idêntica, a utilização de padrões de movimento podem ser importantes a esta etapa, mas não são usuais. Na última etapa, de processamento sequencial dirigido, ocorrem as estratégias visuais de procura apoiadas nas imagens presentes na memória visual, como em uma pesquisa de caminho entre dois símbolos visuais de um mapa a pesquisa visual procura contornos que ligam esses pontos na cor que representam estradas (WARE, 2004).

Outro fator da percepção visual a ser levado em conta é a experiência passada. Para RENSINK (2002) ela é fundamental para o processo da percepção, pois sem a consciência prévia não é possível realizar associações que possibilitem a compreensão. Para além disso, o mesmo autor aponta que na medida que adquirimos novas informações a nossa percepção se altera e isso demonstra, portanto, que a percepção visual é o resultado da interação intrínseca entre informações externas adquiridas pelo sistema visual e informações internas baseadas no conhecimento previamente adquirido.

2.2.2 Comunicação

Considerando os mecanismos genéricos de seleção que todos possuímos (SKINNER, 1972), para que a comunicação desejada não seja interpretada como algo a ser descartado, no caso da visual, podem ser utilizadas técnicas de *design* que favorecem esta situação. Sem almejar aprofundamento na área, entendendo que, conforme DONDIS (1991), um elemento de comunicação visual é constituído de partes, um grupo de unidades determinadas por outras unidades, cujo significado, em conjunto, é uma função do significado das partes, foram buscadas considerações sobre estilos e cores que pudessem contribuir com cada uma das unidades que possam compor o corpo de dados dos símbolos gráficos utilizados para a transmissão das mensagens desejadas.

Os símbolos existem dos mais diversos tipos e, aprendidos da mesma forma que se aprende uma língua, podem identificar direções, ações, estados de espírito. A criação de soluções visuais devem ser regidas pelo significado pretendido, considerando o meio em si, estilo pessoal e cultural (DONDIS, 1991). Outro aspecto visual que possui, por si só, informação é a cor. Sua percepção é o mais emocional dos elementos específicos do processo visual, tem grande força e pode ser usada para expressar e intensificar a informação visual. No meio ambiente, por exemplo, compartilhamos significados associativos da cor das árvores, do céu, da terra e de coisas nas quais vemos as cores como estímulos comuns a todos. DONDIS (1991) afirma que a tudo associamos um significado ou a uma vasta categoria de significados simbólicos e afirma que a cor vermelha, por exemplo, carrega

alguns significados:

> "(...) vermelho, por exemplo, significa algo, mesmo quando não tem nenhuma ligação com o ambiente. O vermelho que associamos à raiva passou também para a "bandeira (ou capa) vermelha que se agita diante do touro". O vermelho pouco significa para o touro, que não tem sensibilidade para a cor e só é sensível ao movimento da bandeira ou capa. Vermelho significa perigo, amor, calor e vida, e talvez mais uma centena de coisas. Cada uma das cores também tem inúmeros significados associativos e simbólicos." (DONDIS, 1991, p. 38).

Dessa forma, a cor pode contribuir para a transmissão de mensagens antes mesmo da identificação de outras unidades do corpo de dados do elemento visual. A Tabela 2 apresenta a relação que algumas cores tem com sensações, como efeitos de distância, temperatura e disposição psíquica. Este autor descreve que a cor vermelha como muito irritante e intranquilizante, podendo estimular o estado de alerta.

Tabela 2: Efeitos psicológicos das cores

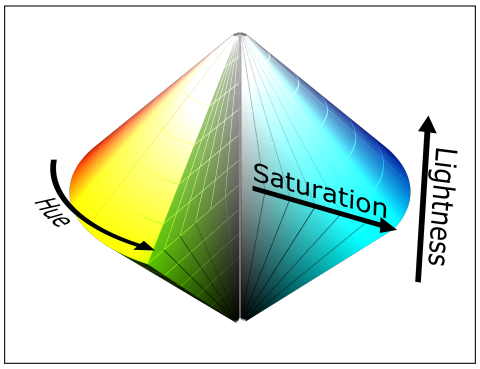
Cor	Efeito de distância	Efeito de temperatura	Disposição psíquica
Azul	Distância	Frio	Tranquilizante
Verde	Distância	Frio a neutro	Muito tranquilizante
Vermelho	Próximo	Quente	Muito irritante e intranquilizante
Laranja	Muito próximo	Muito quente	Estimulante
Amarelo	Próximo	Muito quente	Estimulante
Marrom	Muito próximo, Contenção	Neutro	Estimulante
Violeta	Muito próximo	Muito próximo	Agressivo, intranquilizante, desestimulante.

Fonte: GRANDJEAN (1998)

DONDIS (1991) afirma que existem muitas teorias da cor, que a cor, tanto da luz quanto do pigmento, tem um comportamento único. Ele explica que o conhecimento da cor na comunicação visual vai muito pouco além da coleta de observações das reações a ela. Não há um sistema unificado e definitivo de como se relacionam os matizes. A cor tem três dimensões que podem ser definidas e medidas. Matiz ou croma, é a cor em si, e existe em número superior a cem. Cada matiz tem características individuais; os grupos ou categorias de cores compartilham efeitos comuns.

Em meio digital, existem diversos formatos de composição e representação de característica de cores, como os formatos RGB, que gera as cores na combinação das cores vermelha (*red*), verde (*green*) e azul (*blue*) e HSL, que tem a escolha da matriz (*hue* ou tonalidade) e configuração das propriedades de saturação (*saturation*) e brilho (*lightness*) em unidades percentuais de quantidade de cinza e de branco, respectivamente. Neste

Figura 5: Representação geométrica do espaço de cor HSL



Fonte: Wikimedia Commons - ⟨https://commons.wikimedia.org/wiki/File:HSL_color_solid_dblcone.png⟩.

padrão, como representado na Figura 5, a escolha da cor é baseada no um ângulo do círculo de cores (ou seja, o arco-íris representado em um círculo) onde, por definição, vermelho = 0 = 360, e as outras cores estão espalhadas ao redor do círculo, então verde = 120, azul = 240, etc. A saturação e a luminosidade são representadas como porcentagens, onde 100% é saturação total e 0% é um tom de cinza. Para qualquer matriz onde a luminosidade é 0% a cor é preta e 100% de luminosidade é branca, sendo 50% de luminosidade representa a cor "normal" (W3C, 2018).

Para DONDIS (1991), das três matrizes primárias ou elementares existentes, vermelho, amarelo e azul, a vermelha é a mais ativa e emocional associada com as outras, pode obter novos significados, como abrandar-se e mistura com o azul ou intensificar-se em mistura com o amarelo. Outro tipo de propriedade é a saturação (grau de pureza), que quanto mais saturada for a coloração de um objeto ou acontecimento visual, maior a qualidade de um matiz de cor pelo grau de mesclagem do matiz com a cor branca, que DONDIS (1991) aborda como mais intenso e carregado de expressão e emoção. A terceira propriedade é luminosidade (claridade), que é uma das modalidades tríplices que são usadas para referenciar as cores de um espaço de cores RGB de um modo mais intuitivo e compreensível ao ser humano.

Como a percepção da cor é o mais emocional dos elementos específicos do processo visual, ela tem grande força e pode ser usada com muito proveito para expressar e intensificar a informação visual. A cor não apenas tem um significado universalmente compartilhado através da experiência, como também um valor informativo específico, que se dá através dos significados simbólicos a ela vinculados (DONDIS, 1991).

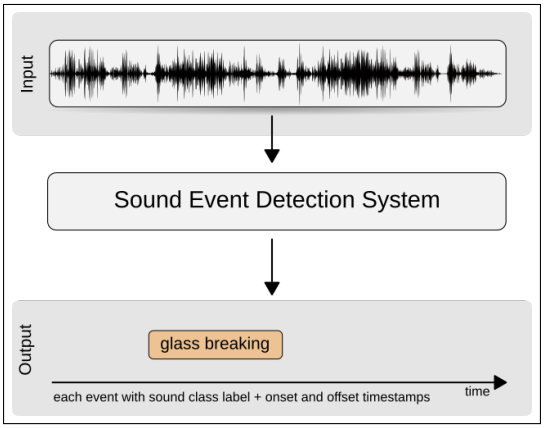
O mesmo autor, afirma que a utilização de simbologia é "extremamente eficaz em termos de comunicação", está impregnado de informação de significado universal, não existe apenas na linguagem. A abstração voltada para o simbolismo requer uma simplificação

radical, ou seja, a redução do detalhe visual a seu mínimo irredutível. O símbolo deve ser simples, referir-se a um grupo, ideia, atividade comercial, instituição, etc. condensar informação, de tal modo que ela possa ser registrada e comunicada. Para ser eficaz, um símbolo não deve apenas ser visto e reconhecido; deve também ser lembrado, e mesmo reproduzido. Não pode, por definição, conter grande quantidade de informação pormenorizada, mas, a abstração pode tornar necessária alguma educação por parte do público quanto ao seu significado.

2.3 Detecção e Classificação de Eventos Acústicos - AED

Os sons trazem uma grande quantidade de informações sobre nosso ambiente cotidiano e eventos físicos que ocorrem nele. Os seres humanos são muito habilidosos em perceber as características gerais da cena sonora ao seu redor, seja uma rua movimentada, um parque silencioso ou um ambiente de escritório silencioso, e reconhecendo fontes de som individuais nas cenas, como carros que passam, pássaros ou passos (MESAROS et al., 2017). Da mesma forma que do sistema auditivo humano combinado com o cognitivo, algumas tecnologias, como representa a Figura 6, buscam soluções para este tipo de habilidade.

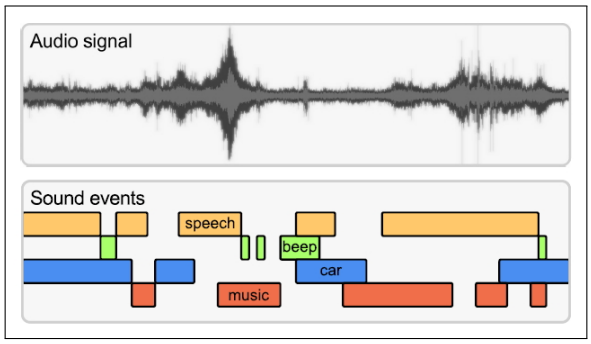
Figura 6: Detecção de eventos sonoros



Fonte: MESAROS et al. (2017)

A Detecção de Eventos Acústicos (AED), também conhecida como detecção de eventos sonoros (*Sound Event Detection* - SED), é definida como reconhecimento de sons individuais em áudio envolvendo também estimativa de início e deslocamento para instâncias distintas de eventos de som. Este tipo de estudo era chamado de detecção monofônica, conforme Figura 7, mas passou a ser chamada de detecção polifônica, pois um som dificilmente ocorre sozinho.

Figura 7: Classificação múltipla de eventos sonoros



Fonte: VIRTANEN (2016)

Uma quantidade significativa de pesquisa ainda é necessária para reconhecer de forma confiável cenas de som e fontes de som individuais em paisagens sonoras da vida real, onde vários sons estão presentes, muitas vezes simultaneamente e distorcidos pelo ambiente (MESAROS et al., 2017) e isto é um dos desafios de AED. Outro desafio é que os sons pertencentes a uma mesma classe podem ter várias fontes, por exemplo, o latido de um cachorro pode ser produzido a partir de várias raças de cães com diferentes características acústicas (CAKIR E.; HEITTOLA, 2015).

Abordagens mais tradicionais nesta área, as quais esta seção é dedicada, utilizam métodos de processamento de linguagem natural (PLN)[3], como *Mel Frequency Cepstral Coefficient* (MFCC) com *Gaussian Mixture Model* (GMM's) combinado com *Hidden Markov Model* (HMM's) (MESAROS A.; HEITTOLA, 2010; HEITTOLA T.; MESAROS, 2013). Outra forma é utilizar técnicas de processamento de imagem, por exemplo, por meio de transformações generalizadas de Hough[4] (HT) como em DENNIS J.; TRAN (2013), para analisar espectrogramas.

Estudos mais recentes obtém melhores resultados utilizando *Feedforward Neural Networks* (FNN's)[5], na forma de *Multi Layer Perceptrons* (MLP's)[6] utilizando características de espectrogramas gerados a partir de eventos sonoros sobrepostos temporalmente em áudios de ambientes realistas (CAKIR E.; HEITTOLA, 2015). Recentemente, é notável o surgimento de métodos que utilizam aprendizado profundo em AED, com muitos sistemas surgindo sendo baseados em vários tipos de redes neurais profundas (DNNs)

[3]O Processamento de Linguagem Natural (PLN) estuda a capacidade e as limitações de uma máquina em entender a linguagem dos seres humanos.

[4]A transformada de Hough é um método para detecção de formas que são facilmente parametrizadas (linhas, círculos, elipses, etc.) em imagens computacionais. Mais informações na seção 2.3.4.

[5]Feedforward Neural Networks são redes neurais onde a saída de uma camada é usada como entrada para a próxima camada, ou seja, não há loops na rede.

[6]A perceptron multicamadas (MLP) é uma rede composta por mais de uma camada de neurônios ligadas entre si por sinapses com pesos.

(MESAROS et al., 2017).

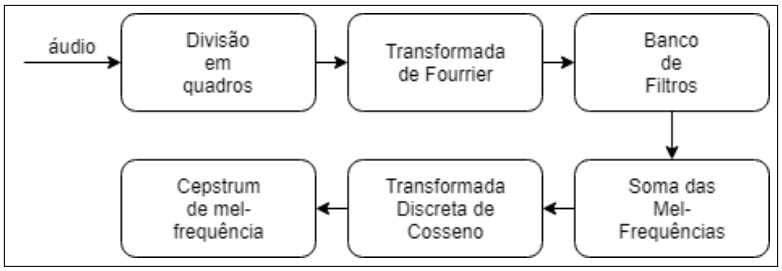
2.3.1 *Mel Frequency Cepstral Coefficient* - MFCC

A extração da melhor representação paramétrica dos sinais acústicos é uma tarefa importante para produzir um melhor desempenho de reconhecimento (MUDA, 2010). A eficiência dessa tarefa, em fase de pré-processamento, é importante, pois, afeta o comportamento e desempenho das próximas fases. O cepstrum de mel-frequência (MFCC) é uma representação do espectro de potência de curto prazo de um som, baseado em uma transformação de cosseno linear de um espectro de potência de registro em uma escala de mel não linear de frequência, uma das técnicas de extração de características mais populares usadas, por exemplo, no reconhecimento de fala (RAZAK Z.; IBRAHIM, 2008; HASAN M. R.; JAMIL, 2004).

O MFCC é baseado em percepções de audição humana que não conseguem perceber frequências acima de 1Khz. Em outras palavras, o MFCC é baseado na variação conhecida da largura de banda crítica do ouvido humano em função da frequência. Ele possui dois tipos de filtro que são espaçados linearmente em baixa frequência abaixo de 1000 Hz e espaçamento logarítmico acima de 1000 Hz e sua realização consistem em cinco etapas: pré-processamento, enquadramento, janelas, DFT, Mel Filterbank, Logaritmo e Inverse DFT (MUDA, 2010).

Cada etapa da extração da MFCC realiza os seguintes procedimentos: enquadramento do sinal em quadros curtos, cálculo de estimativa do periodograma do espectro de potência de cada quadro, aplicação de banco de filtros no espectro de potência e soma da energia em cada filtro, é tomado o logaritmo de todas essas energias, realizada a transformada discreta de cosseno (DCT) das energias do banco de filtros de log e mantidos os coeficientes DCT, o restante é descartado. Essas etapas, representadas na Figura 8, são detalhadas abaixo.

Figura 8: Extração do *Mel Frequency Cepstral Coefficients*

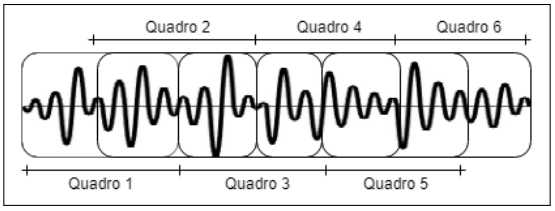


Fonte: Adaptado de HASAN M. R.; JAMIL (2004)

Na primeira etapa o sinal é dividido em quadros de alguns milissegundos, por exemplo, um quadro de 25 ms em um sinal de 16kHz tem o comprimento de 0,025*16000 =

400 amostras. Um novo quadro geralmente tem seu início em 10 ms após o início do primeiro quadro (160 amostras depois), ou seja, como pode ser observado em Figura 9, os quadros tem algum trecho dos sinais que se sobrepõem e isso acontece até o final do arquivo.

Figura 9: Divisão de arquivo de áudio em quadros



Fonte: autor

Na segunda etapa é reduzida a distorção espectral aplicando uma função de janela amortecida, um código de bloco linear como o de Hamming[7]. Já na terceira etapa é aplicada a Transformada Rápida de Fourier (FFT) na janela para mostrar a magnitude, obtendo o espectro. Uma FFT é um algoritmo que converte um sinal num período de tempo (ou espaço), por exemplo um arquivo WAV, para uma representação no domínio da frequência. O espectro de frequência de um sinal é a distribuição das amplitudes e fases de cada componente de frequência em relação à frequência.

Na quarta etapa o sinal é passado para a escala de Mel, uma escala psicoacústica de tons de sons, no sentido de sua identificação entre baixo e agudos, cuja unidade é mel. Está relacionada à Hertz (Hz), a unidade de medida do Sistema Internacional de Frequências, por uma relação baseada na audição humana. É uma escala de frequência mais próxima do que o ouvido humano é capaz de captar. A fórmula de transferência é, onde m a escala Mel e f a frequência, a seguinte:

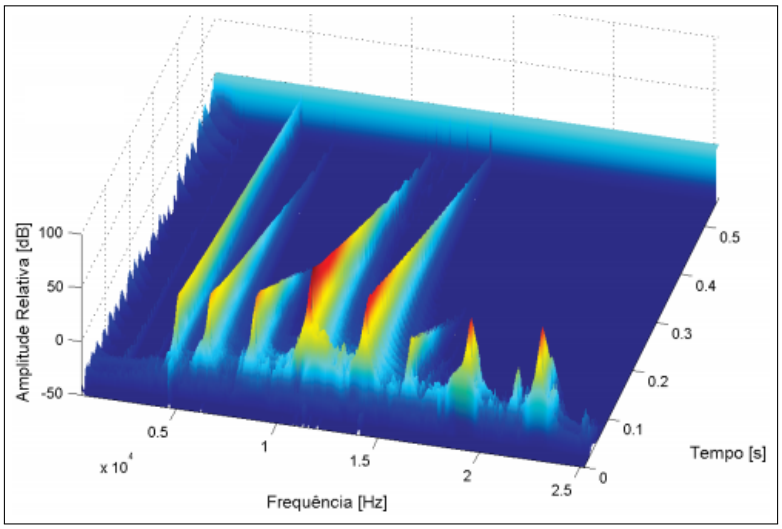
$m = 2595 \log (1 + f / 700)$

Ao final é convertido o espectro logarítmico da escala de Mel em escala de tempo usando a transformada discreta de cosseno e é obtido o cepstrum de mel-frequência, ou MFCC. Um cepstrum é o resultado da tomada da transformação inversa do logaritmo do espectro estimado de um sinal.

Dessa forma, a extração de características de áudios e geração de espectrogramas com essas informações, em suma, transformam os sinais de áudio em sinais visuais, que podem ser representados por gráficos do tipo cascata no domínio tempo-frequência, como apresentado no em Figura 10. Este gráfico permite visualizar a variação temporal das am-

[7]O código de Hamming é um código de bloco linear, desenvolvido por Richard Hamming, é utilizado no processamento de sinal e nas telecomunicações. A sua utilização permite a transferência e armazenamento de dados de forma segura e eficiente.

Figura 10: Representação temporal-frequencial de som



Fonte: Análise Acústica - ⟨http://www.espiraldotempo.com/oldsite/wp-content/uploads/2013/02/ET35_11_JLC_Acustica.pdf⟩.

plitudes das componentes frequenciais do som em análise. Em gráficos planos a energia sonora é representada unicamente pela coloração, sendo o vermelho a representação de maior intensidade. Na Figura 10, por exemplo, é possível observar que a maior parte da energia sonora existe aproximadamente a partir dos 4 kHz.

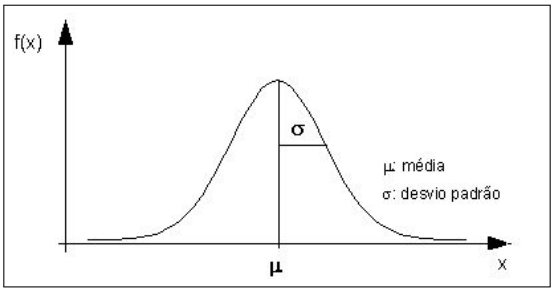
2.3.2 *Gaussian Mixture Model* - GMM

A distribuição gaussiana, também conhecida como a distribuição normal, é um modelo amplamente utilizado em probabilidade e estatística para a distribuição de variáveis contínuas, para a modelagem de fenômenos naturais. Surge em muitos contextos diferentes e pode ser motivada a partir de uma variedade de diferentes perspectivas. Uma situação na qual a distribuição gaussiana surge é quando consideramos a soma de múltiplas variáveis aleatórias. O teorema do limite central (devido a Laplace) nos diz que, sujeito a certas condições suaves, a soma de um conjunto de variáveis aleatórias, que é uma variável aleatória, tem uma distribuição que se torna cada vez mais gaussiana como o membro de termos a soma aumenta (BISHOP, 2006).

A Figura 11 representa o gráfico da função gaussiana, curva de Gauss ou curva em forma de sino, a distribuição normal, correspondente à equação 1 utilizada para definir a curva gaussiana de uma dimensão onde x é um conjunto com n valores, tal que $-\infty < x < \infty$, G representa a distribuição gaussiana dos valores de x, σ o desvio padrão dos valores de x, referindo-se à dispersão da distribuição, tal que $\sigma > 0$, e μ a média dos valores de x, indicando a posição central da distribuição.

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(\mu-x)^2}{2\sigma^2}} \quad (1)$$

Figura 11: Gráfico de distribuição normal padrão (curva de Gauss ou curva em forma de sino).



Fonte: Adaptado de BISHOP (2006)

Apesar da distribuição Gaussiana ter importantes propriedades analíticas, ela é limitada na modelagem de situações reais quando, por exemplo, os dados possuem dois grupos dominantes. Nestes casos, não é possível capturar essa estrutura com apenas uma distribuição Gaussiana, mas, visando fornecer uma classe de modelos de densidade mais rica do que a única Gaussiana, pode ser formulado um modelo probabilístico de mistura de distribuições mais básicas, como uma superposição linear simples de componentes Gaussianos, criando um modelo de misturas. Em geral, os modelos de mistura podem compreender combinações lineares de outras distribuições, no caso abordado nesta seção, o modelo abordado é de mistura Gaussianas, chamado *Gaussian Mixture Model* (BISHOP, 2006).

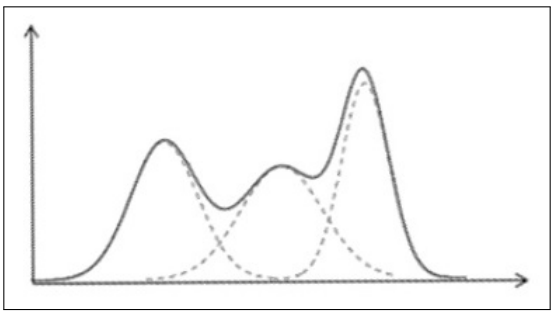
A Figura 12 apresenta um modelo de mistura gaussiana onde é possível observar a combinação linear de três gaussianos, cada um dimensionado por um coeficiente (em pontilhado), dando origem a uma média de covariâncias (linha contínua). Uma combinação linear de gaussianos, onde f_1 é a densidade de $N(\mu_1, \sigma_1^2)$ e f_2 a densidade de $N(\mu_2, \sigma_2^2)$, então $\lambda f_1 + (1-\lambda) f_2$, pode dar origem a muitas médias e covariâncias, bem como aos coeficientes na combinação linear, quase qualquer densidade contínua pode ser aproximada a uma precisão arbitrária (BISHOP, 2006).

2.3.3 *Hidden Markov Model* - HMM

Uma cadeia de Markov, chamada assim em homenagem ao matemático Andrei Andreyevich Markov, é um caso de processo estocástico[8] com estados discretos (o

[8]Dentro da teoria das probabilidades, um processo estocástico é uma família de variáveis aleatórias representando a evolução de um sistema de valores com o tempo. É a contraparte probabilística de um processo determinístico. Ao invés de um processo que possui um único modo de evoluir, como nas soluções

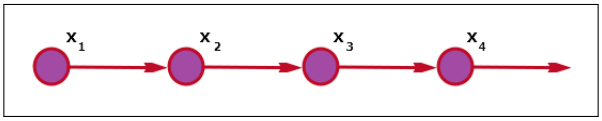
Figura 12: Exemplo de um modelo de mistura Gaussianas em representação de uma dimensão.



Fonte: BISHOP (2006)

parâmetro, por exemplo o tempo, pode ser discreto ou contínuo) com a propriedade de que a predição do próximo estado depende apenas do estado atual, conforme apresenta a Figura 13, e não na sequência de eventos que precederam, uma propriedade chamada de Markoviana. Cadeias de Markov têm aplicações como modelos estatísticos de processos que seguem uma cadeia de eventos ligados do mundo real, como no processamento de linguagem natural (MESAROS et al., 2017).

Figura 13: Cadeia de Markov de primeira ordem.



Fonte: BISHOP (2006)

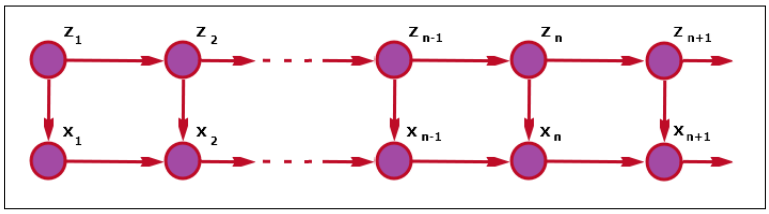
Da mesma forma que um sistema dinâmico linear, como o modelos GMM apresentado na seção anterior (2.3.2), uma cadeia de Markov pode ser adequadamente caracterizada usando a estrutura de modelos gráficos probabilísticos, podendo ser descrito, como na Figura 14, por grafos direcionados com uma estrutura de árvore (sem loops) para os quais a inferência pode ser executada de forma eficiente usando o algoritmo soma-produto[9]. Um HMM, que pode ser construído a partir de componentes mais simples, é como uma

de equações diferenciais ordinárias, por exemplo, em um processo estocástico há uma indeterminação: mesmo que se conheça a condição inicial, existem várias, por vezes infinitas, direções nas quais o processo pode evoluir.

[9]O algoritmo de soma-produto, também conhecido como propagação de crença, é um algoritmo de passagem de mensagens para realizar inferências em modelos gráficos, como redes bayesianas e campos aleatórios de Markov. Calcula a distribuição marginal para cada nó (ou variável) não observado, condicionada a quaisquer nós (ou variáveis) observados.

instância específica do modelo de espaço de estados no qual as variáveis latentes são discretas. Dessa forma, para cada observação Xn há uma variável latente correspondente Zn, satisfazendo a propriedade de independência condicional de chave que Zn-1 e Zn+1 são independentes, dado Zn. (BISHOP, 2006).

Figura 14: Cadeia de Markov representada por um modelo de espaço de estados.



Fonte: BISHOP (2006)

2.3.4 Transformada de Hough - HT

Para um ponto (x_i, y_i) a equação da linha que o contém pode ser representada por: $y_i = ax_i + b$. Existem linhas infinitas que atendem esta equação para vários fatores de a (declive) e b (ordenada para origem). Examinando o espaço paramétrico de a e b (plano ab) pode ser visto que um ponto (a_i, b_i) representa uma certa linha no plano xy, ou seja, $b = - x_i a + y_i$. Dessa forma, para cada ponto (x_i, y_i) pertencendo a uma linha, outra linha é obtida no espaço paramétrico ab. Todas essas linhas se cruzam em um ponto (a', b'). Este ponto é a representação da linha no plano xy que contém todos os pontos (x_i, y_i), colineares (que têm declive a' e são ordenados para a origem b'). O problema que surge dessa transformação é para os casos em que a linha no plano ab é vertical, pois a inclinação da linha é infinita. Para resolver este problema, norte americano Paul Hough patenteou em 1962 um método alternativo que consiste em expressar a equação da linha em coordenadas polares, a chamada Transformada de Houg (HT). (CANTO, 2012).

A equação básica da transformada de Hough foi projetada determinar os parâmetros de objetos geométricos simples, que possam ser descritas de forma paramétrica, como linhas e círculos, presentes em imagens computacionais, onde cada ponto da imagem deve ser representado por uma reta ou por uma senoide (MASEK et al., 2003). Isso significa que qualquer linha reta no espaço da imagem xy é representada por um único ponto no espaço de parâmetros $\rho\theta$, e qualquer parte desta linha reta é transformada no mesmo ponto. $\rho = x_i cos(\theta) + y_i sen(\theta)$. O alcance do ângulo θ varia de $-\pi/2$ a $\pi/2$ medido em relação ao eixo x. E o intervalo de ρ varia de $-N\sqrt{2}$ a $N\sqrt{2}$, onde N é a dimensão da imagem quadrada. O método de Hough subdivide o espaço de parâmetros nas chamadas células acumuladoras. A célula coordenada i, com valor acumulador A_{ij}, corresponde ao quadrado associado às coordenadas (CANTO, 2012).

Utilizada em visão computacional, sua aplicação até meados da década de 80 estava sendo muito lenta devido à complexidade computacional de armazenamento e a dificuldade de um entendimento detalhado de suas propriedades. Contudo, nos últimos anos, muitas pesquisas têm sido realizadas, contribuindo assim para um melhor aproveitamento de seus recursos (MARRONI, 2002) e, atualmente, por meio de transformações generalizadas pode ser utilizado na análise de espectrogramas para resolução de problemas de AED (CAKIR E.; HEITTOLA, 2015).

Entretanto, há vários problemas com o método de transformação Hough. Primeiro de tudo, requer que os valores limite sejam escolhidos para a detecção de bordas, e isso pode resultar na remoção de pontos de borda críticos, resultando na falha na detecção de um arco de círculo. Em segundo lugar, a transformação de Hough é computacionalmente intensiva devido à sua abordagem de "força bruta" e, portanto, pode não ser adequada para aplicações em tempo de execução (MASEK et al., 2003).

2.3.5 Redes Neurais Artificiais

Redes Neurais Artificiais são algoritmos de Aprendizado de Máquina (*Machine Learning* - ML) compostos por técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência, uma abordagem recente para soluções em AED (CAKIR E.; HEITTOLA, 2015). Aprendizagem de Máquina, por sua vez, é uma das áreas de estudo da Inteligência Artificial (IA) e esta, uma subárea da Ciência da Computação relacionada ao estudo de agentes racionais, de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas. O Aprendizado de Máquina é uma das áreas de estudo em IA, que visa desenvolver sistemas para realizar tarefas que são melhor realizadas por seres humanos que por máquinas, ou não possuem solução algorítmica viável pela computação convencional.

As Redes Neurais Artificiais (ANN) são modelos computacionais que possuem neurônios com pesos e possibilidade de configuração por meio de aprendizado. Esses modelos, inspirados em redes neurais biológicas, são organizados em camadas e quando recebem uma entrada (isto é, um único vetor) transformam a entrada através das camadas por meio de conjuntos de neurônios, onde cada neurônio recebe várias entradas e realiza uma soma ponderada delas, seguindo para uma função de ativação, e responde com uma saída. Um padrão frequentemente utilizados é o de conexão completa inter-camadas (*full connected*), apenas na direção entrada-saída (*feed-forward*) e nenhuma conexão intra-camada (THOMÉ, 2002; FADLULLAH Z. M.; TANG, 2017).

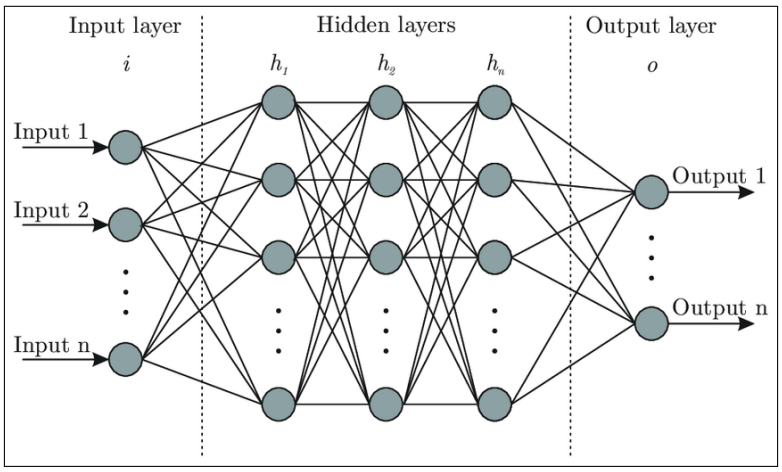
A função que define o neurônio pode ser vista na equação 2, onde $a^{\int}$ representa o nível de resposta (ou valor de ativação) para as unidades na camada $\int$ (a_i^l a ativação de i em $\int$), W é uma matriz de peso, onde $W_{ij}^{\int}$ representa o parâmetro (ou peso) associado à conexão entre a unidade j na camada $\int$ e a unidade i na camada $\int$+1. $b^{\int}$ é o viés, também

chamado de bias, associado às unidades da camada $\int$ e σ (desvio padrão) representa a função de ativação (ou não-linearidade). A saída da rede é definida pelas ativações das unidades na camada mais profunda (ORDONEZ F. J.; ROGGEN, 2016).

$$\mathbf{a}^{(l+1)} = \sigma\left(W^l \mathbf{a}^l + \mathbf{b}^l\right) \tag{2}$$

A Função Linear Retificada (*Retified Linear Unit* - ReLU) é um exemplo de função de ativação. Definida por $f(x) = max(0, x)$, é uma função não linear de ativação de neurônios. A principal vantagem de usar a função ReLU sobre outras funções de ativação é que ela não ativa todos os neurônios ao mesmo tempo, pois se a entrada do neurônio for negativa, ela será convertida em zero e o neurônio não será ativado (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 15: Rede Neural Artificial - ANN



Fonte: FADLULLAH Z. M.; TANG (2017)

No modelo de uma rede neural *feed-forward* tradicional, um neurônio situado em determinada camada tem sua saída conectada com todos os neurônios da camada seguinte e a nenhum outro neurônio de camadas anteriores, posteriores ou sua própria (Figura 15) (THOMÉ, 2002). Ao final, na camada de saída, são comumente utilizadas funções *Softmax* para gerar a representação probabilística de cada classe para cada valor de entrada.

Na função *Softmax*, conforme definido na equação 3, os resultados só podem assumir valores entre 0 e 1, e a soma da probabilidade de todas as classes é igual a 1. Desta forma é possível determinar que a classe estimada pela rede é a que possuir maior probabilidade (GOODFELLOW; BENGIO; COURVILLE, 2016).

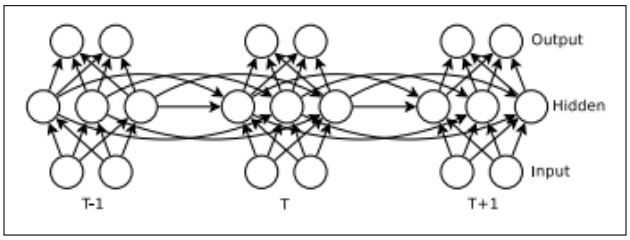
$$softmax(z_i) = \frac{exp(z_i)}{\sum_j exp(z_j)} \tag{3}$$

Em fase de treinamento, para otimização da rede, é frequentemente utilizado o método do Gradiente Descendente (ou *Gradient Descent* - GD). O GD é um algoritmo para encontrar o mínimo de uma função, utilizado como método de otimização para minimizar o erro da rede neural. Esta minimização é realizada modificando os pesos e os limites de ativação com o objetivo de encontrar o mínimo local da função perda.

Outro tipo de estrutura que compõe o universo de modelos de redes neurais é do tipo recorrente. Uma Rede Neural Recorrente (RNN) possui realimentação, onde um neurônio pode ser direta ou indiretamente retroalimentado pela sua saída. Cada camada pode conter conexões entre os elementos de processamento da mesma camada (estímulos laterais), das camadas anteriores e das camadas posteriores (THOMÉ, 2002).

Na topologia recorrente não existe um sentido único para o fluxo de sinais entre neurônios ou entre camadas. Essas redes podem ter sua arquitetura considerada generativa profunda (SUTSKEVER, 2013). A profundidade de um RNN pode ser tão grande quanto o comprimento da sequência de dados de entrada (Figura 16). E, por conta dessa característica, o RNN é particularmente útil para modelar os dados da sequência em eventos sonoros HOCHREITER (1998).

Figura 16: Rede Neural Recorrente - RNN



Fonte: SUTSKEVER (2013)

Um problema destes tipos de rede *full connected*, que recebe como entrada tipicamente um vetor e a estrutura tem conectividade total entre as camadas é sua dimensionalidade. O aumento do número de camadas ocasiona o aumento da complexidade e do tempo de processamento da rede. Aumentando o número de neurônios por camada, por exemplo, quando a entrada se torna muito grande e complexa, como o treinamento de imagens de alta resolução (FADLULLAH Z. M.; TANG, 2017), acarreta o aumento do grau de liberdade da função de transferência, e quanto maior a quantidade de variáveis livres, menor será a capacidade de generalização da rede (THOMÉ, 2002).

Afora o problema da dimensionalidade, o uso de RNNs foi restrito até recentemente devido ao chamado problema de fuga/dissipação do gradiente (*vanishing problem*). O Problema da Dissipação do Gradiente ou *The Vanishing Gradient Problem*, que também pode ocorrer em ANNs, é um fenômeno que pode ocorrer em fase de treinamento com

método do Gradiente Descendente, por conta dos neurônios de camadas anteriores aprenderem muito mais lentamente que os neurônios das camadas posteriores. Com isso o gradiente tende a diminuir à medida que nos movemos para trás através das camadas ocultas, tendendo a explodir ou a desaparecer. Para solucionar problemas como de dimensionalidade em ANNs foram propostas camadas de convolução em vez de conectividade total nas camadas da rede neural. Para o problema de dissipação de gradiente aparecem na literatura métodos de otimização para treinar RNNs geradores que modificam a descida de gradiente estocástica e modelos recorrentes de aprendizagem profunda, como Redes de Memória de Longo Prazo (LSTMs) (MIKOLOV T.; KARAFIáT, 2010; SUTSKEVER, 2013). Estes modelos, abordados na próxima seção (2.3.6), dão origem às Redes Neurais Profundas ou *Deep Learning* (DL).

2.3.6 Redes Neurais Profundas

Nos últimos anos, a pesquisa em aprendizado profundo ganhou um impulso notável tanto na academia quanto na indústria. O aprendizado profundo é uma nova geração da técnica de Aprendizado de Máquina, que está ganhando muita popularidade e ampla utilização em vários campos da ciência da computação, como reconhecimento de objetos, reconhecimento de voz, processamento de sinais, robótica, jogos de AI e assim por diante (SEIDE F.; LI, 2011; YU D.; SELTZER, 2013; FADLULLAH Z. M.; TANG, 2017).

As Redes Neurais Profundas (*Deep Learning*), da mesma forma que as ANNs, são aplicadas a computadores inspiradas na forma de funcionamento dos cérebros (UTGOFF P. E.; STRACUZZI, 2002; BENGIO Y.; LECUN, 2007). A organização de conceitos ocorre de forma hierárquica, do conceito mais simples à combinação de conceitos para representações abstratas e para isso, utilizam de uma quantidade de dados maior, com menor número de camadas em proporção. Estes modelos também se diferem em como as camadas são combinadas e seus respectivos papéis. Por exemplo, em redes convolucionais (um modelo de aprendizagem profunda) uma camada tem o objetivo de modelar um padrão específico, que pode ser encaminhado para camadas sucessivas, enquanto em um modelo de ANN uma camada tem o papel de projetar os dados em um subespaço onde o problema é linearmente separável (FADLULLAH Z. M.; TANG, 2017).

Nas técnicas tradicionais de Aprendizado de Máquina, conforme pode ser observado na Figura 17, a maioria dos recursos aplicados precisa ser identificada por um especialista em domínio para reduzir a complexidade dos dados e tornar os padrões mais visíveis para os algoritmos de aprendizado funcionarem. A maior vantagem dos algoritmos do Deep Learning é que eles tentam aprender recursos de alto nível a partir de dados de maneira incremental. Diferente das redes tradicionais, essa capacitada das redes profundas elimina a necessidade de conhecimento de domínio e profunda extração de recursos (HUSSAIN et al., 2019).

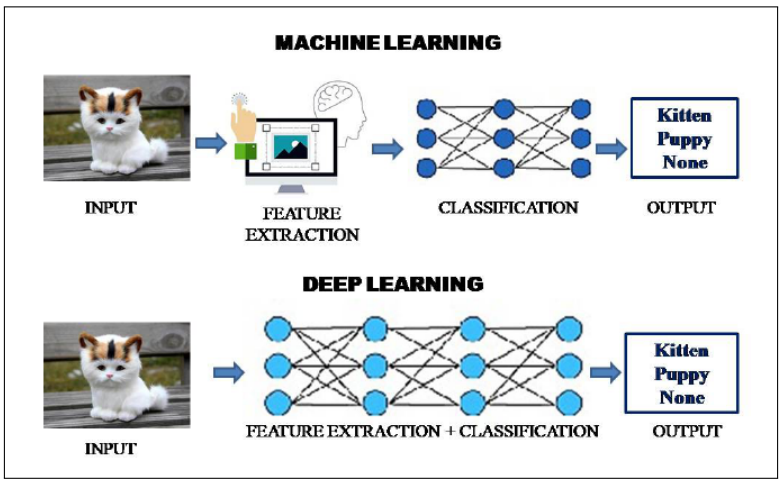
Outra grande diferença entre a técnica Deep Learning e Machine Learning é a abor-

dagem de resolução de problemas. As técnicas de Aprendizagem Profunda tendem a resolver o problema de ponta a ponta, onde as técnicas de aprendizado de Máquina precisam que as declarações de problemas se dividam em diferentes partes a serem resolvidas primeiro e depois seus resultados serão combinados no estágio final (HUSSAIN et al., 2019).

Por exemplo, para um problema de detecção de múltiplos objetos, técnicas de AP podem fazer uso da imagem como entrada e a rede fornece a localização e o nome dos objetos na saída, diferente de algoritmos de ML, onde é necessário primeiro identificar todos os objetos possíveis e estas saídas serem enviadas para outra rede para a classificação. Neste sentido, onde por um lado a necessidade de execução de procedimentos em separado pode dificultar todo o processo, por outro, como em algoritmos de árvore de decisão e regressão logística, dá melhor interpretabilidade ao processo, pois possibilita análise de respostas para etapas específicas. Em DL se pode descobrir quais nós de uma rede foram ativados, mas não o que eles deveriam estar modelando e qual a função destas camadas em relação ao coletivo (DENG, 2014).

Já em relação ao desempenho um algoritmo de *Deep Learning* pode levar muito tempo para treinar em comparação com a mesma abordagem em ML, isso ocorre devido ao grande número de parâmetros. O cenário se inverte na fase de testes, pois em ML o tempo de teste aumenta com o aumento do tamanho dos dados (DENG, 2014).

Figura 17: A extração de características em ML e DL



Fonte: HUSSAIN et al. (2019)

Dependendo de como as arquiteturas são planejadas, as redes neurais profundas podem ser categorizadas em três tipos: aprendizado não supervisionado ou generativo, aprendizagem supervisionada e híbridas.

Uma arquitetura de aprendizado não supervisionado visa caracterizar as proprieda-

des de correlação de alta ordem dos dados de entrada para fins de análise ou síntese de padrões quando não há informações sobre os rótulos das classes na base. Por outro lado, uma arquitetura para aprendizagem supervisionada é usada para fins de classificação ou reconhecimento de padrões, tem os rótulos sempre disponíveis e pode ser chamada de rede profunda discriminativa. Combinando as arquiteturas generativa e discriminativa, um modelo híbrido pode ser construído para fazer tarefas de discriminação, onde sua discriminação é assistida com os resultados de redes profundas generativas ou não supervisionadas (BENGIO, 2009; PANG Y.; SUN, 2017; DENG, 2014).

Além dos esforços de pesquisa na academia, pesquisadores baseados no setor também estão dedicando um grande esforço para aplicações de aprendizagem profunda para reconhecimento de voz e processamento de sinais.

O sistema de voz do MAVIS (*Microsoft Audio Video Indexing Service*)[10] baseado em aprendizagem profunda, por exemplo, mostrou uma queda significativa na taxa de erro em comparação com as técnicas contemporâneas de ML (por exemplo, misturas gaussianas para a modelagem acústica) (SEIDE, 2019), porém, os vetores de entrada usados na rede são de dimensão fixa, podendo não ser adequada para o reconhecimento de sequência de fala quando a dimensionalidade das entradas e/ou saídas pode ser variável.

Por outro lado, o HMM, baseado em operações de programação dinâmica, é útil para modelar dados de sequência de eventos com comprimento variável. Portanto, para explorar suas respectivas vantagens, os pesquisadores consideraram o uso conjunto do classificador estático (isto é, a rede neural profunda) e o HMM (SUTSKEVER, 2013). Além disso, LSTMs e CNNs também foram utilizados para lidar com o problema de dimensionalidade mencionado acima, envolvendo as entradas e / ou saídas (WESTON J. R.; RATLE, 2012; SOCHER R.; PERELYGIN, 2013; SOCHER R.; HUANG, 2011). Estes métodos foram aplicados a vários conjuntos de dados musicais, e os resultados demonstraram uma melhoria do erro relativo de 5% a 30% em comparação com o método de transcrição polifônica existente (BENGIO Y.; BOULANGER-LEWANDOWSKI, 2013; RIFAI S.; BENGIO, 2012).

Conforme GOODFELLOW; BENGIO; COURVILLE (2016), o que difere aprendizado de máquina de um problema de otimização é necessidade de técnicas capazes de treinar modelos que generalizam exemplos nunca antes processados. Ou seja, não é suficiente que o modelo tenha bom desempenho apenas na base de treinamento, ele necessita passar por uma validação do treinamento. O treinamento consiste no ajuste dos pesos sinápticos e vieses dos neurônios de modo que o vetor de saída se aproxime da saída esperada.

O processo de treinamento ocorre em duas etapas principais: a propagação

[10] O MAVIS é um serviço de indexação de áudio e vídeo da Microsoft que utiliza tecnologia de reconhecimento de voz desenvolvida na Microsoft Research para permitir a pesquisa de arquivos de áudio e vídeo com fala. Mais informações em ⟨https://www.microsoft.com/en-us/research/project/mavis/⟩.

(*feedfoward-propagation*) e a retro-propagação (*back-propagation*). A primeira segue a camada de entrada, passa pelas camadas ocultas e termina na camada de saída, onde a rede entrega os valores estimados. Estes valores de saída são comparados com o desejado e a função perda é definida, estabelecendo então o desempenho da rede. Se o desempenho não foi suficiente é iniciada a fase de retro propagação, onde se deseja minimizar o erro da estimação, por exemplo, utilizando o método de Gradiente Descendente.

Para a realização deste processo de treinamento é necessário ter dois conjuntos de dados, divididos em treinamento e teste, com o objetivo de diminuir tanto o erro de treinamento quanto o de teste. Quando estes dois erros não caminham juntos, ocorre um problema importante chamado sobre-ajuste *overfitting*. O *overfitting* ocorre quando o erro de treinamento decresce, mas o de teste continua alto. Ou seja, quando a rede se ajusta tão bem ao conjunto de dados que acaba se mostrando ineficaz para prever entradas diferentes das utilizadas para treinamento. Outro problema, semelhante a este, é o *underfitting*, quando o erro de treinamento não reduz, ou seja, a rede acaba não aprendendo tudo o que pode com os arquivos de treinamento, não foi capaz de determinar uma relação entre os dados (GOODFELLOW; BENGIO; COURVILLE, 2016).

Em fase de treinamento, um dos algoritmos de otimização em aprendizagem profunda mais utilizados visando solucionar o problema fuga/dissipação do gradiente, abordado como um dos problemas das ANNs (seção 2.3.5), é o método do Gradiente Descendente Estocástico (SGD) (GOODFELLOW; BENGIO; COURVILLE, 2016). Uma adaptação do método Gradiente Descendente utilizado em ANNs, o SGD, projetado para atuar em grandes conjuntos de dados, visa economizar o custo computacional de cada iteração, calculando uma estimativa da perda, utilizando uma pequena parte de treinamento.

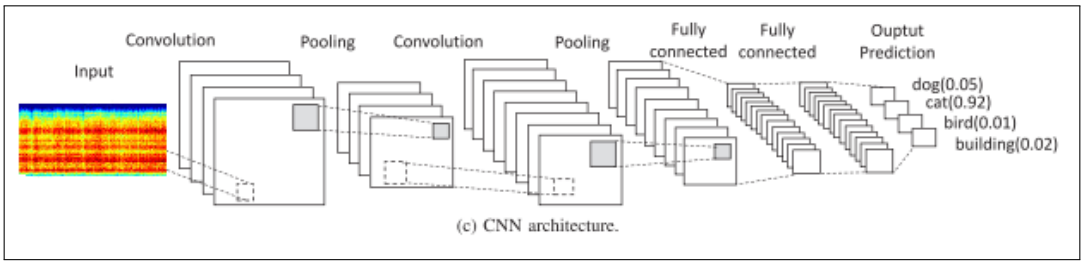
2.3.7 Rede Neural Convolucional - CNN

Uma Rede Neural Convolucional, também conhecida como Convnet, é uma arquitetura de rede profunda discriminativa que, possui neurônios com pesos e possibilidade de configuração por meio de aprendizado, como as ANNs (FADLULLAH Z. M.; TANG, 2017).

A arquitetura de uma CNN para lidar com uma entrada complexa consiste em várias camadas de convoluções, que podem ser temporais (1D, de uma dimensão, usado em processamento de fala e entendimento de linguagem natural), espacial (2D, para segmentação e classificação de imagens) ou volumétrica (3D, análise de vídeos) com funções de ativação não-linear para calcular a saída. Por conta disso, a CNN compreende conexões, como mostrado na Figura 18, localizadas em que cada região da entrada é conectada a um neurônio na saída. Cada camada aplica filtros diferentes (na ordem de centenas a milhares) e combina seus resultados. Além disso, a CNN consiste nas camadas de agrupamento para subamostragem (FADLULLAH Z. M.; TANG, 2017).

Este tipo de arquitetura é particularmente adaptada para classificar imagens e tenta

Figura 18: Rede Neural Convolucional - CNN



Fonte: Adaptado de FADLULLAH Z. M.; TANG (2017)

tirar proveito, para além da análise local entre pixels próximos, da estrutura espacial das imagens. O uso dessa arquitetura torna as redes convolucionais rápidas de treinar. Isso, por sua vez, ajuda a treinar redes profundas de muitas camadas (FADLULLAH Z. M.; TANG, 2017).

Durante a fase de treinamento, uma CNN aprende automaticamente os valores de seus filtros com base na tarefa determinada. Por exemplo, para classificar eventos sonoros a partir de imagens, assuma que os pixels brutos de uma imagem que representa um espectrograma compõem a entrada de uma CNN (Figura 18). Na primeira camada, a CNN pode aprender a detectar as arestas dos pixels brutos, na segunda empregar as arestas para detectar formas e nas camadas subsequentes, utilizando estas formas ter a capacidade de aprender características de nível superior, características de cada tipo de evento. Na camada final, um classificador é usado para explorar esses recursos de alto nível (FADLULLAH Z. M.; TANG, 2017).

Geralmente, as CNNs são treinadas empregando métodos de aprendizado supervisionados, nos quais um grande número de pares de entrada-saída é essencial. No entanto, obter um conjunto de treinamento substancialmente grande tem sido um desafio na aplicação de CNNs para resolver novas tarefas. As CNNs mostraram-se bem sucedidas no aprendizado para solução de tarefas específicas, obtendo resultados melhores, principalmente em tarefas de visão computacional, que técnicas contemporâneas de ML (FADLULLAH Z. M.; TANG, 2017).

2.3.8 Memória Longa de Curto Prazo - LSTM

O modelo de rede Memória Longa de Curto Prazo (*Long Short-Term Memory* - LSTM) foi criado em 1997, por Hochreiter e Schimdhuber, para resolver a limitação que ocorre em métodos de aprendizagem baseados em gradientes, como propagação retroativa através do tempo e aprendizagem recorrente em tempo real. Em fase de treinamento, a evolução temporal da integral do caminho sobre todos os sinais de erro que ”retornam no tempo”depende exponencialmente da magnitude dos pesos e isso implica que o erro

retro propagado rapidamente desaparece ou explode. Portanto, os RNNs não aprendem na presença de intervalos de tempo maiores que 5 - 10 etapas de tempo discretos entre eventos de entrada relevantes e sinais de destino (HOCHREITER; SCHMIDHUBER, 1997).

Porém esta estrutura pode falhar em aprender a processar corretamente certas séries temporais muito longas ou contínuas que não são a priori segmentadas em subsequências de treinamento apropriadas com princípios e fins claramente definidos. O problema é que um fluxo de entrada contínuo pode eventualmente fazer com que os valores internos das células cresçam sem limite, mesmo que a natureza repetitiva do problema sugira que eles sejam redefinidos ocasionalmente. Este problema é abordado em GERS; SCHMIDHUBER; CUMMINS (1999) com a proposta de adição de uma porta de esquecimento para resolver este problema.

Por conseguinte, uma LSTM padrão é composta de uma célula, uma porta de entrada, uma porta de saída e uma porta de esquecimento. A célula, impondo erros constantes através de Carrosséis de Erro Constante (CECs), implementa a capacidade da rede aprender a corrigir atrasos mínimos em mais de 1000 etapas de tempo discreto. As três portas, (entrada, saída e esquecimento) são unidades multiplicativas e aprendem a regular o fluxo de informações para dentro e para fora da célula (GERS; SCHMIDHUBER; CUMMINS, 1999).

2.3.9 Artigos Relacionados

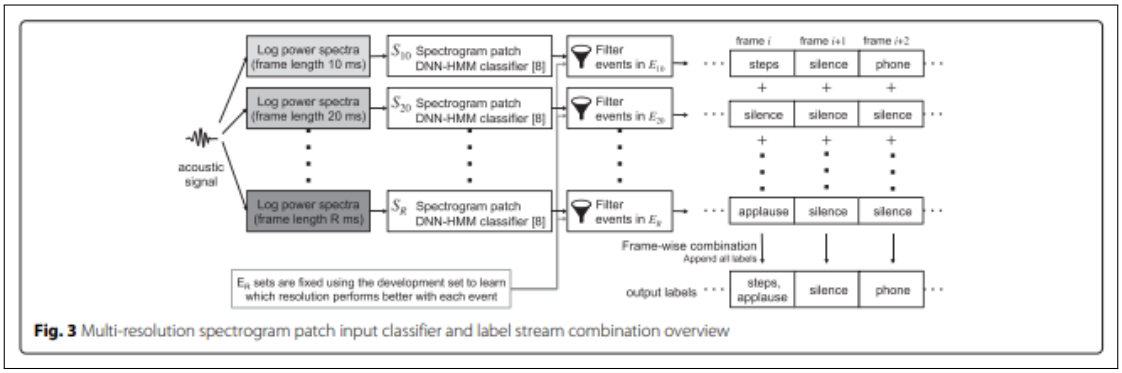
Apoiado pela colocação de CAKIR E.; HEITTOLA (2015), que recentemente arquiteturas de CNNs e RNNs apresentaram performance melhores que métodos estáveis tradicionais, utilizam uma, outra ou essas duas, formando uma rede neural convolucional recorrente (CRNN), arquiteturas combinadas que, com suas soluções e *datasets* distintos, provem soluções de AED, dando apoio em termos de tecnologia ao desenvolvimento de uma proposta de solução relacionada ao problema motivador desta pesquisa.

Em ESPI M.; FUJIMOTO (2015), buscando o reconhecimento de eventos acústicos, como aplausos, tosse, movimento de cadeira etc., em arquivos de gravações de fala em seminários são utilizados métodos de pré-processamento para converter o som em um espectrograma de alta definição, uma imagem, e submeter a rede neural. O autor critica estudos que utilizam como entrada da rede espectrograma gerados a partir de divisões dos arquivos de som, afirma que dessa forma são perdidas propriedades do som, então, conduz seu estudo comparando a performance de uma rede neural profunda e uma rede neural convolucional que recebem como entrada uma pilha de quadros de diferentes resoluções (tempo e frequência) que passaram pela transformada de Fourier[11]. São forne-

[11] A transformada de Fourier é uma transformada integral que expressa uma função em termos de funções de base sinusoidal. A transformada de Fourier, epônimo a Jean-Baptiste Joseph Fourier, decompõe uma função temporal (um sinal) em frequências.

cidos rótulos de saída em paralelo, que são sintetizados, mesclando rótulos repetidos ou removendo rótulos de silêncio caso exista algum rótulo conhecido em mesmo espaço de tempo (Figura 19).

Figura 19: Esquema de funcionamento de rede CNN multi-resolução de espectrograma



Fonte: ESPI M.; FUJIMOTO (2015)

Os resultados utilizando arquitetura de redes neurais convolucionais apresentaram desempenho melhor que arquitetura de rede neural profunda tradicional. O esquema de combinação da abordagem traz como contribuição um modelo híbrido de aprendizagem que não utiliza recursos que focam nas propriedades específicas de determinados sons. Isto sugere que o desenvolvimento desta pesquisa utilize de redes neurais convolucionais.

O trabalho de PARASCANDOLO G.; HUTTUNEN (2016), visando mapear ocorrência de eventos sonoros (por exemplo, música, carro, fala) em arquivos da vida real de 10 contextos cotidianos diferentes, também realiza a conversão do sinal sonoro em espectrogramas para a solução de AED. O método utilizado normaliza as amplitudes, divide em frames de 50 milissegundos com 50% de sobreposição, como na Figura 9, e diminui a magnitude do espectro para 40 bandas. Depois, cada áudio se torna um longo vetor com uma sequência de características. Cada vetor original é dividido de três formas em sequências menores, resultando em vetores que correspondem a segmentos de 0.25, 0.62 e 2.5 segundos do áudio original.

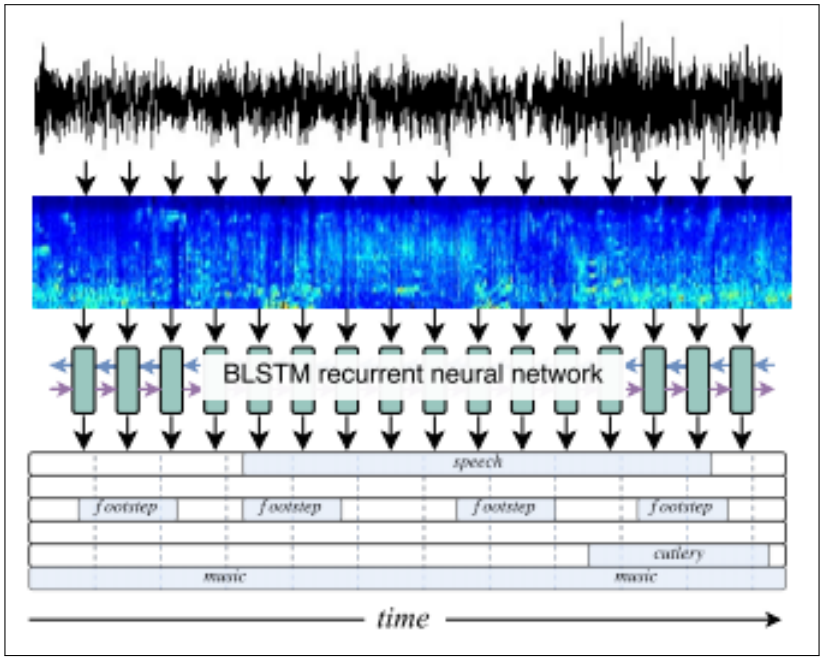
O *dataset* deste trabalho teve áudios de 44.1 kHz e 24-bit, convertidos para mono, os dois canais em um, pelo método de cálculo de médias de Krylov-Bogolyubov *averaging*, um método matemático para análise aproximada de processos oscilantes em mecânica não linear. O método baseia-se no princípio da média, quando a equação diferencial exata do movimento é substituída pela sua versão média e foi utilizado na etapa de treinamento, na tentativa de reduzir o *overfitting*.

Com 60 classes, o *dataset* teve seus arquivos aumentados por transformações simples nos espectrogramas (simulações de aceleração ou desaceleração, deslocamento de tempo

de subquadro, blocos de mistura) adicionado ruído gaussiano, um ruído estatístico cuja função densidade de probabilidade (FDP) é igual a da distribuição normal, que é também conhecida como distribuição gaussiana. Nos áudios, uma classe a mais foi adicionada para representar evento desconhecido.

Dividido em 60% dos arquivos para treinamento, 20% para validação e 20% para teste. Foi treinado com o método de *early stopping*, que permite especificar um grande número arbitrário de épocas de treinamento e interromper o treinamento quando o desempenho do modelo parar de melhorar em um conjunto de dados de validação de espera. Cada segmento foi submetido a uma rede neural recorrente RNN BLSTM[12] com várias camadas ocultas onde cada frame é associado a um vetor de classes de áudio, indicando se alguma classe está presente ali ou não (Figura 20).

Figura 20: Detecção de Eventos Sonoros com arquitetura RNN



Fonte: PARASCANDOLO G.; HUTTUNEN (2016)

O autor destaca que que os RNNs têm cerca de 850K parâmetros cada, em comparação com os parâmetros 1,65M da FNN treinados com os mesmos dados. Desse modo, os RNNs fazem um uso mais eficiente e efetivo dos parâmetros, devido às conexões recorrentes e à estrutura mais profunda com camadas menores, portanto, em comparação com FNN's os resultados não foram muito diferentes, mas RNN se saíram melhores, com menos parâmetros, mais eficientes e efetivos (PARASCANDOLO G.; HUTTUNEN, 2016).

[12]BLSTM (*Bidirectional Long Short-Term Memory Networks*), algoritmo utilizado em redes neurais que tem a capacidade de acessar o contexto de longo alcance, aprender o alinhamento de sequências e trabalhar sem a necessidade de dados segmentados

Estes resultados apontam a possibilidade de maior eficiência para identificar eventos de alerta utilizando uma arquitetura de rede com camadas recorrentes.

O estudo de CAKIR E.; HEITTOLA (2015) afirma que CNNs são boas em extrair características dos sons e RNNs em aprender o contexto temporal dos áudios e recentemente essas arquiteturas apresentaram performance melhores que métodos estáveis tradicionais. Sua proposta combina esses dois tipos de rede para a detecção de eventos em sons da vida real, formando uma estrutura de rede chamada *Convolutional Recurrent Neural Network*, por possuir camadas convolucionais e recorrentes, e aplica em detecção de eventos sonoros em áudios polifônicos que independem de cena. Sua proposta apresentou melhora de performance em comparação com redes CNNs, RNNs e outros métodos estabelecidos.

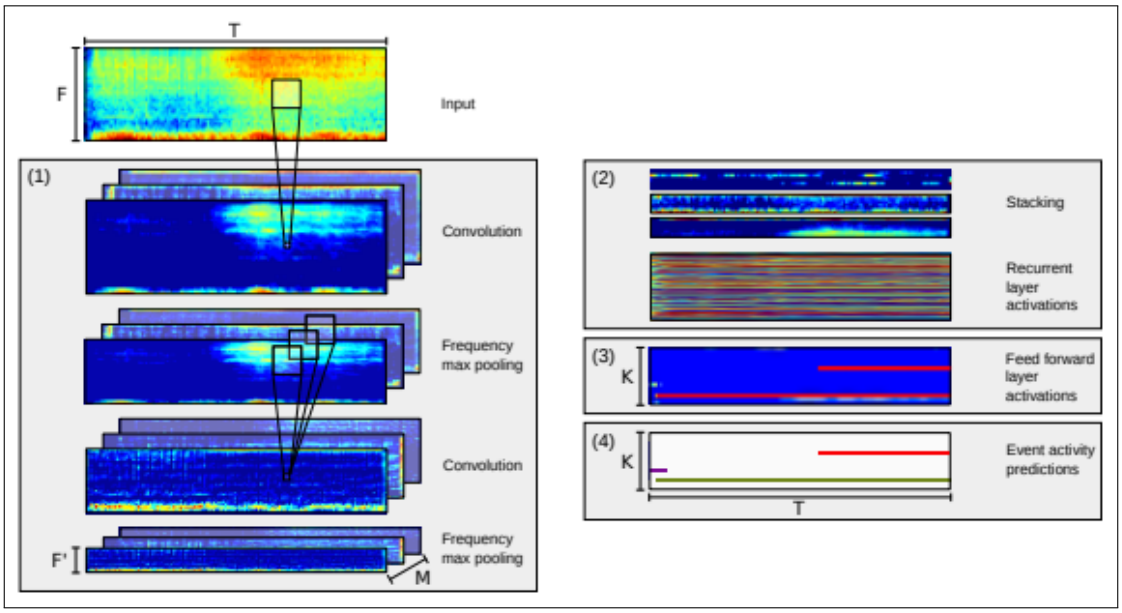
O autor aponta duas grandes deficiências nas arquiteturas de redes neurais profundas para solucionar problemas de AED: invariância de tempo e frequência, o que permitiria modelar pequenas variações nos eventos; e o contexto temporal é restrito a janelas de tempo curto, impedindo a modelagem efetiva de eventos tipicamente mais longos (por exemplo, chuva) e correlações de eventos (CAKIR E.; HEITTOLA, 2015). As CNNs podem abordar a primeira limitação aprendendo filtros que são deslocados no tempo e na frequência, como em ESPI M.; FUJIMOTO (2015), faltando, no entanto, informações de contexto temporal mais longas. As RNNs resolvem a última falha integrando informações das janelas de tempo anteriores, apresentando uma informação de contexto teoricamente ilimitada, como em PARASCANDOLO G.; HUTTUNEN (2016).

No entanto, os RNNs não capturam facilmente a invariância no domínio da frequência, tornando mais difícil a modelagem de alto nível dos dados. Para se beneficiar de ambas as abordagens, as duas arquiteturas podem ser combinadas em uma única rede com camadas convolucionais seguidas por camadas recorrentes, da mesma forma que ocorreram em abordagens recentes para problemas de reconhecimento de fala e classificação musical (PARASCANDOLO G.; HUTTUNEN, 2016).

A modelagem da rede CRNN proposta, representada na Figura 21, consiste em quatro partes: (1) no topo da arquitetura, uma representação de tempo de frequência dos dados (uma janela de contexto de F log mel band energias sobre T frames) é alimentada para camadas convolucionais $L_c \in \mathbb{N}$ com agrupamento não sobreposto sobre o eixo de frequência; (2) os mapas de características da última camada convolucional são empilhados sobre o eixo da frequência e alimentados para camadas recorrentes $L_r \in \mathbb{N}$; (3) uma única camada de *feedforward* com ativação sigmóide lê as saídas da camada recorrente final e estima probabilidades de atividade de evento para cada quadro e (4) probabilidades de atividade de evento são binarizadas por limiar sobre uma constante para obter previsões de atividade de evento (CAKIR E.; HEITTOLA, 2015).

Nessa estrutura, as camadas convolucionais atuam como extratores de características, as camadas recorrentes integram as características extraídas ao longo do tempo, fornecendo assim informações de contexto e, finalmente, a camada de *feedforward* produz as

Figura 21: Arquitetura de rede CRNN



Fonte: Adaptado de CAKIR E.; HEITTOLA (2015)

probabilidades de atividade para cada classe. Esta arquitetura utiliza dos dois modelos propostos nos artigos anteriormente expostos nesta sessão e apresenta, na utilização de camadas de rede tanto de CNN como de RNN, uma proposta de arquitetura que pode obter melhores resultados na detecção de eventos sonoros.

2.3.10 Modelo de Rede Neural para AED

No campo deste projeto, um modelo recente de rede para AED, por exemplo, é formada por duas etapas: representação e classificação sonora. Na etapa de representação, detalhado na seção 2.3.1, são extraídas as características do áudio por meio de uma transformação por MFCC, para cada intervalo de tempo t no sinal de áudio para obter um vetor de feixe de luz $x_t \in \mathbb{R}^F$, onde F $\in \mathbb{N}$ é o número de informações por quadro (CAKIR E.; HEITTOLA, 2015) e no estágio de classificação são utilizadas arquiteturas de redes neurais profundas, como as Redes Neurais Convolucionais (CNNs) e LSTMs, as quais são abordadas nas seções 2.3.7 e 2.3.8, respectivamente.

No estágio de classificação, por exemplo utilizando uma CNN, quando ela recebe uma entrada, são estimadas as probabilidades $\rho(y_t(k)|x_t, \Theta)$ para classes de eventos k = 1, 2,. . . , K no quadro t, onde Θ representa os parâmetros do classificador. As probabilidades de ativação do evento são então binarizadas pelo limiar, por exemplo uma constante, para obter as predições de evento $\hat{y}_t \in \mathbb{R}^K$. Os parâmetros do classificador Θ são treinados por aprendizado supervisionado, e o setor de saídas de destino y_t, com informações de cada

quadro, são obtidas a partir das anotações de início/deslocamento das classes de eventos de som. Se a classe k estiver presente durante o quadro t, y_t (k) será definido como 1, e 0 caso contrário. Como o vetor de saída pode ter múltiplos elementos diferentes de zero, as classes que são localizadas em período de tempo consecutivo são mescladas para representar a saída da rede. Este modelo foi usado para prever a atividade das classes de eventos de som quando as anotações de início/deslocamento não estão disponíveis, como em situações da vida real (CAKIR E.; HEITTOLA, 2015).

Ainda sobre formas de classificação, analisando o problema de certos eventos que não puderem ser facilmente distinguidos por características impulsivas, como um vidro quebrando, enquanto alguns eventos sonoros tipicamente continuam por um longo período de tempo (por exemplo, um bebê chorando), métodos de classificação que podem preservar o contexto temporal ao longo dos vetores de características sequenciais devem ser considerados para esses problemas. Colaborando com a solução destes casos podem ser utilizadas RNNs, nela os recursos de entrada são apresentados como uma matriz de contexto $X_{t:t+T-1}$, onde T $\in \mathbb{N}$ é o número de quadros que define o comprimento de sequência do contexto temporal e a matriz de saída de destino $Y_{t:t+T-1}$ é composta das saídas y_t dos quadros t para t + T - 1 (CAKIR E.; HEITTOLA, 2015).

2.3.11 Métricas de Avaliação de Resultados

As métricas de avaliação de resultados de pesquisas envolvendo redes neurais variam dependendo do tipo de problema de aprendizado de máquina que está sendo resolvido e, essencialmente, analisam o erro entre o vetor de previsões e o vetor de efetivos. A seguir serão apresentadas algumas métricas, apontadas por MESAROS A.; HEITTOLA (2016), utilizadas para problemas de detecção e classificação de eventos sonoros.

2.3.11.1 Precisão, Recall e F1 Score

Uma medida de desempenho deve ter em conta as previsões sobre todo o conjunto de instâncias e, então, que para problemas de classificação bipartida, onde cada arquivo de entrada poderá receber a classificação de quaisquer das múltiplas classes configuradas na rede, são necessárias funções de perda adicionais (SHALEV-SHWARTZ S.; BEN-DAVID, 2014). As funções de perda dependem dos 4 números a seguir:

- TN ou Verdadeiro Negativo: previsão e resultado foram negativas
- TP ou Verdadeiro Positivo: previsão e resultado positivos
- FN ou Falso Negativo: previsão negativa, mas resultado positivo
- FP ou Falso Positivo: previsão positiva e resultado negativo

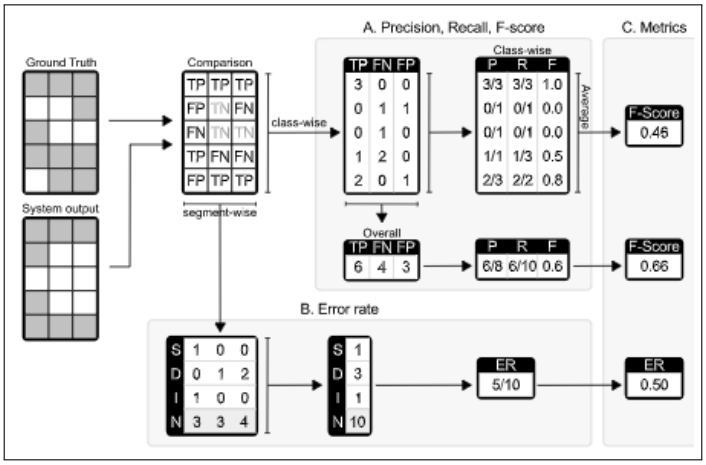
A precisão (*recallprecision*), representada pela letra "P" na equação 4, é calculada pelo número de predições corretas, positivos verdadeiros (classificados como pertencentes

a uma classe que realmente são daquela classe) dividido pela soma entre este número, e o número de falsos positivos (classificados nesta classe, mas que pertencem a outras).Já o *recall*, letra "R" na mesma figura, representa o número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe, dividido pela quantidade total de exemplos que pertencem a esta classe, mesmo que sejam classificados em outra. No caso binário, positivos verdadeiros divididos por total de positivos.

$$P = \frac{TP}{TP + FP'} \quad R = \frac{TP}{TP + FN} \tag{4}$$

O *F1 Score* é uma combinação que forma a média harmônica da precisão e recuperação das predições, muito boa quando você possui um dataset com classes desproporcionais. Tem valor máximo 1 quando a precisão e a recuperação são 1, e valor mínimo 0 sempre que um deles é 0. A vantagem de usar essa métrica para avaliar o desempenho da detecção de eventos sonoros é que ele é amplamente conhecido e fácil de entender, para este tipo de solução pode ser calculado de duas maneiras, baseado em segmento ou baseado em eventos.

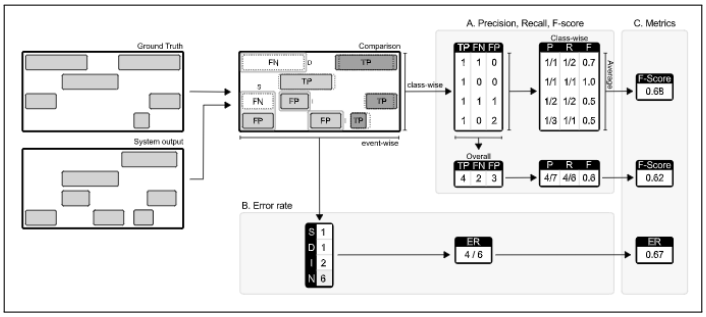
Figura 22: *F1 Score* baseado em segmento



Fonte: SHALEV-SHWARTZ S.; BEN-DAVID (2014)

Os P, R e F baseados em segmento são calculados com base nas estatísticas intermediárias baseadas em segmento, usando a média baseada em instância ou a média baseada em classe (Figura 22). Para o cálculo baseado em eventos são calculados da mesma forma, conforme ilustrado na Figura 23, porém, a partir das estatísticas intermediárias baseadas em eventos (SHALEV-SHWARTZ S.; BEN-DAVID, 2014).

Figura 23: *F1 Score* baseado em eventos



Fonte: SHALEV-SHWARTZ S.; BEN-DAVID (2014)

3 PROCEDIMENTOS METODOLÓGICOS

Este capítulo apresenta três seções de procedimentos práticos, onde a primeira descreve coleta do tipo consulta realizada com pessoas surdas ou que têm alguma relação com a surdez objetivando validar a proposta e identificar características importantes a serem abordadas na solução. A segunda representa o processo de comunicação visual do projeto, com sugestões de utilização de símbolos e suas características. E a terceira descreve a coleta de arquivos importantes para o treinamento de uma rede neural que atenda os propósitos da pesquisa e os experimentos realizados com arquiteturas de rede utilizando estes arquivos.

Os experimentos (seção 4), tiveram o escopo limitado a tecnologias que incorporam as áreas de Processamento de Sinal (*Signal Processing* - SP) e Detecção de Eventos Sonoros (*Sound Event Detection* - SED), representados na Figura 24 nas cores verde e amarelo, respectivamente. A área de PS está presente na aplicação de filtros e transformada MFCC realizadas em etapa de pré-processamento, e a área de detecção de eventos nas tecnologias de redes neurais para aprendizagem e classificação dos sinais processados.

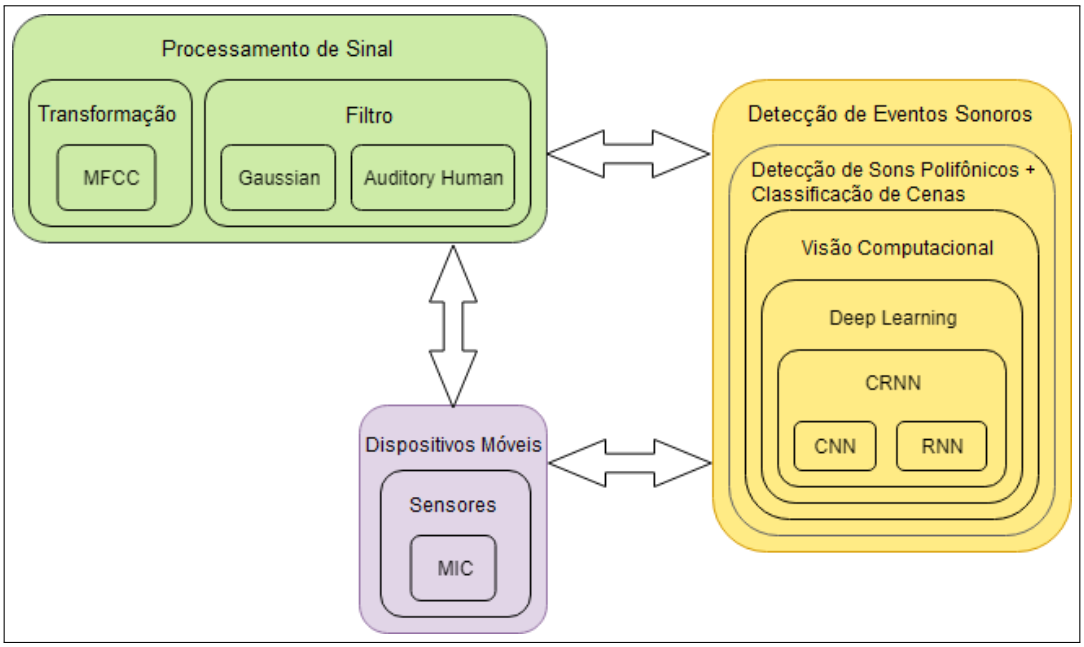
O quadro na cor roxa apresentado (Figura 24) é composto por áreas relacionadas ao desenvolvimento *mobile*, sugerindo que a solução idealizada possa ser utilizada em *smartphones*. Pesquisas e desenvolvimentos abordando a utilização de redes neurais em dispositivos móveis, utilização de sensores disponíveis em *smartphones* (*phone sensing*), como o microfone, são sugeridos como trabalhos futuros em seção específica (seção 5.1).

3.1 Levantamentos Prévios

Para dar maior suporte à idealização de uma solução tecnológica, auxiliar a definição de requisitos do objeto e nortear o processo de desenvolvimento desta pesquisa envolvendo possíveis beneficiários, foi elaborado um instrumento visando o levantamento de dados que apoiem a etapa inicial de projeto. O questionário, conforme demonstrado no Apêndice A, foi submetido a pessoas surdas ou com alguma relação com a surdez.

Elaborado em formato estruturado em recurso online, o questionário de coleta inicial para definição e validação da proposta teve como motivação coletar informações sobre

Figura 24: Áreas de pesquisa relacionadas



Fonte: autor

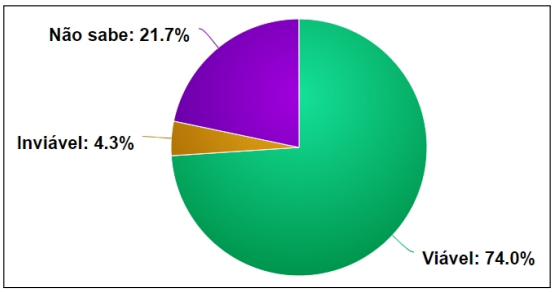
a necessidade dos surdos em perceber de alguma forma os sinais de alerta que ocorrem nos ambientes; verificar se esta funcionalidade contribuiria para proporcionar uma vida independente e inclusão; identificar quais tipos de ocorrências são interessantes que um dispositivo seja capaz de alertar; se alguma tecnologia já é utilizada por eles para este fim; contribuições e críticas acerca deste tema. Como forma de aperfeiçoar a elaboração do instrumento, o questionário passou previamente pela experiência do teste, por meio do qual as perguntas formuladas foram respondidas por um surdo, um ouvinte que tem estudos relacionados com acessibilidade e um ouvinte matriculado em curso de extensão em LIBRAS.

Para execução da coleta o formulário foi enviado para a Feneis e outros contatos do pesquisador que tem relação com a surdez, 23 pessoas responderam, sendo que 11 informaram serem surdas, 10 professores de LIBRAS, 5 intérpretes, 5 tem familiar(es) surdo(s), 1 tem perda auditiva, 1 trabalha com Pessoas com Deficiência, 1 tem relação profissional com surdo e 2 tem relação de amizade.

Questionados sobre a viabilidade da apresentação de alerta visual por meio de reconhecimento de som em ambiente externo ou controlado, conforme pode ser observado na Figura 25, a maioria, dezessete dos respondentes consideram viável, já cinco não sabe responder e um julgou não ser viável uma tecnologia que reconheça sons, representando 73.9%, 21,7% e 4,3%, respectivamente.

Antes mesmo do instrumento apresentar sugestões para tipos de sons de alertas a se-

Figura 25: Viabilidade de reconhecimento de som



Fonte: autor

rem informados aos surdos os respondentes apontaram 7 tipos de sons. Dentre as respostas estão sirenes e alarme de emergência com cinco ocorrências, campainhas com quatro e o restante, buzinas, choros de bebês, acidentes e tumultos com uma ocorrência cada. Dessas, acidentes e tumultos ainda não haviam sido previstas para a proposta.

Ainda sobre dados obtidos na questão de tipos de alertas foram obtidas, também, informações não objetivadas pela pergunta como sinais visuais ou luminosos para forma de exposição das informações, duas citaram algum tipo de vibração e uma telões. Ainda, foram informados lugares onde essa tecnologia seria importante: escolas, residências, empresas, aeroportos e locais públicos.

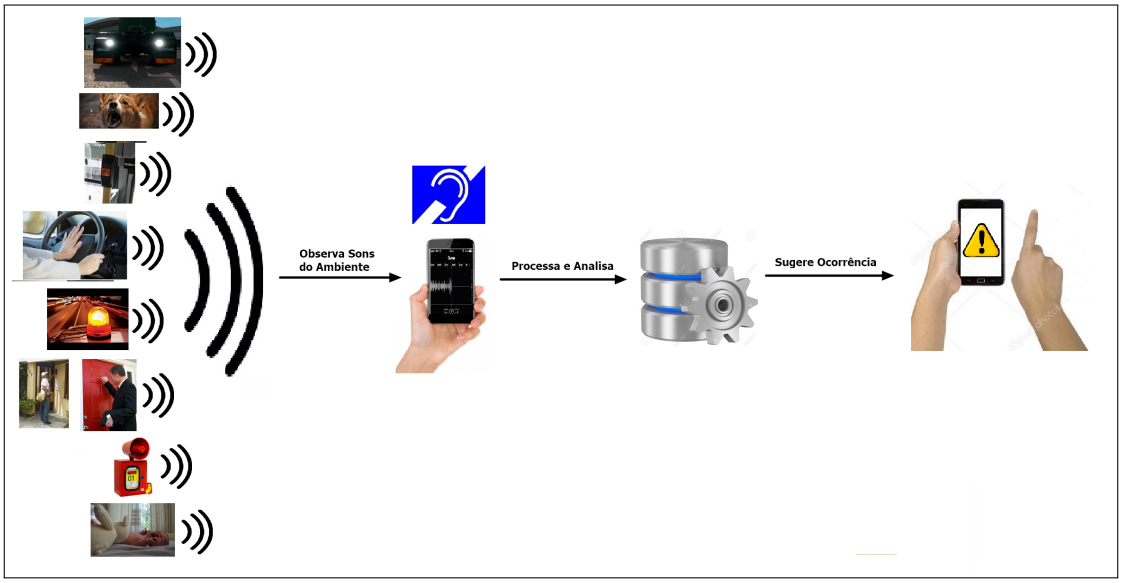
Após, foi apresentada a Figura 26 que ilustra a ideia de funcionamento do projeto. Nela, é apresentado o emissor de sons e uma sequência de imagens simbolizando a forma de processamento e apresentação da mensagem ao usuário.

Alguns aspectos julgados importantes aos respondentes aparecem na imagem de representação do sistema sobre a forma de apresentação da informação final. Uma delas está representada na imagem como alerta visual na tela de um smartphone, a segunda não está representada, mas o smartphone pode ser capaz de vibrar no momento de alerta. Já a terceira indicação de forma de apresentação, por telões, não estaria contemplada na imagem e, diferentemente das demais indicações, não seria possível entrar no escopo do projeto.

Na sequência foi questionado o conhecimento dos respondentes sobre recurso com mesma funcionalidade. As respostas, apresentadas na Figura 27, foram em sua maioria negativas, compreendendo 52,2% das respostas. Outros 30,4% não responderam negativamente, mas também não citaram algum tipo de tecnologia. Já 17,4% responderam conhecer alguma tecnologia, citando babás eletrônicas, campainhas residenciais luminosas. Observa-se que nenhuma das tecnologias citadas nas respostas tem estrutura semelhante à proposta.

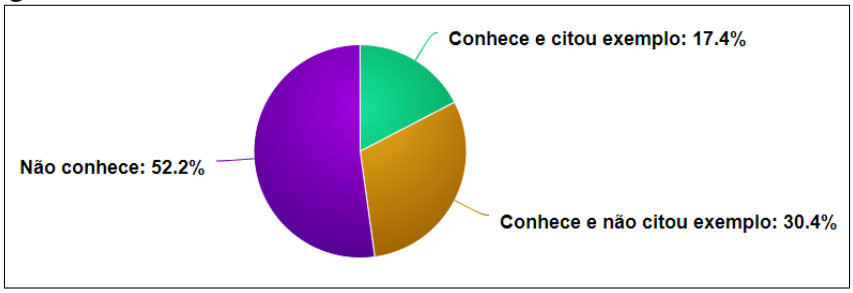
A penúltima questão sobre a contribuição desta tecnologia obteve que, conforme pode

Figura 26: Estrutura conceitual com sugestões de tipos de sons



Fonte: autor

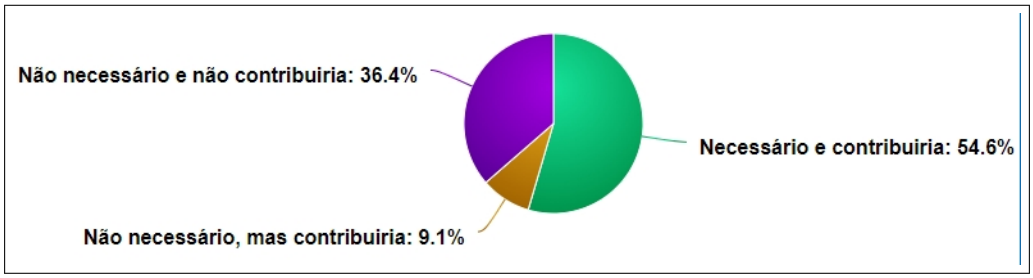
Figura 27: Tem conhecimento de recursos com mesma funcionalidade



Fonte: autor

ser verificado na Figura 28, para 78,3% é uma necessidade, 13% opina o contrário, mas aponta que contribuiria e 8,7% alegaram não ser uma necessidade e não contribuir para vida independente e inclusão de surdos.

Figura 28: Contribuição da proposta para os surdos



Fonte: autor

O último espaço do questionário, para inserção de texto livre, foi dedicado a sugestões e críticas. Nele, alguns respondentes salientaram o interesse da ideia, inclusive para outros graus de deficiência auditiva, pedido para viabilizar o projeto o quanto antes e que seja comum a todos. Um deles sugerindo formas para viabilizar a construção e relatando que o sucesso poderia "(...)acarretar um marco na comunidade surda(...)". Outra sugestão é sobre a importância de envolver as pessoas as quais o projeto pretende atingir: "Tentar colher informações concretas do dia a dia dos surdos com eles próprios pois saberão informar com mais precisão as necessidades e dificuldades que passam em seu cotidiano".

A pesquisa apontou por pessoas que têm relação com pessoas surdas que a utilização de uma tecnologia pode beneficiar a vida de surdos e promover autonomia e isto complementa a motivação para o desenvolvimento da proposta. Validou a proposta e auxiliou no ranqueamento dos tipos de eventos sonoros importantes aos surdos. Com estes resultados foi possível adequar a proposta e definir quais tipos de eventos devem ser priorizados no desenvolvimento da ferramenta, desta forma, o instrumento cumpriu o que se esperou da pesquisa.

3.2 Símbolos Visuais

A primeira característica a ser definida para a sugestão de comunicação visual foi está ligada à cores dos ícones e do fundo onde eventualmente serão apresentados. Principalmente por conta do aspecto emocional que as cores representam ao processo visual, foram definidas as cores, conforme Tabela 3, preta, azul, vermelha e branca. Para informações de alerta foi utilizada predominantemente, como fundo da imagem, cor vermelha, na composição 100% de saturação e 30% de luminosidade, em combinação com a cor branca para melhor contrastar com a anterior, a cor amarela foi evitada em combinação com a

vermelha, pois a interpretação por parte do usuário poderia não corresponder a intenção de apenas alertar o usuário de um possível evento. Em contrapartida a escolha pela cor azul, intenciona informar o usuário apenas que o sistema está em funcionamento e, por conta de suas sensações psíquicas de abrandar, causar tranquilidade, não causar distração.

Tabela 3: Propriedades HSL de cores

Cor	Matriz	Saturação	Luminosidade
Preto	0º	0%	0%
Azul	240º	100%	30%
Vermelho	0º	100%	30%
Branco	0º	0%	100%

Fonte: autor

Como sugestão dos símbolos, foram buscadas imagens observando fatores de pregnância, simplicidade e, visando a possibilidade de experiência passa em outros contextos, que pudessem ser resgatadas, reproduzidas e utilizadas livremente. Desta forma, foi utilizada a coleção de imagens de símbolos *Open Symbols*[1]. A escolha por uma ferramenta gratuita e disponível na *Internet* está na importância da eficácia do símbolos quando são facilmente reconhecidos e podem ser reproduzidos. Esta coleção oferece símbolos diversos, extraídos de várias fontes e ofereceu imagens sugestivas de representação para os tipos de alerta abordados nesta pesquisa. , resultando nas subfiguras apresentadas na Figura 29.

Dos símbolos selecionados, durante a pesquisa foi reconhecido o símbolo utilizado para representação de Dispositivos Auxiliares de Audição (ALDs), este símbolo se tornou importante para a pesquisa, pois, já possui associação com este tipo de instrumentos que ajudam a amplificar os sons auditivos e auxiliam na decodificação do som, podendo, por exemplo, facilitar a identificação, pela pessoa surda, do tipo de recurso ou sua aplicação.

3.3 Elaboração de *Dataset*

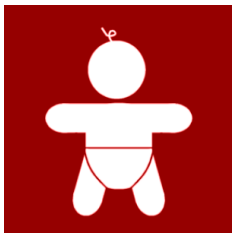
Esta seção denota a busca, seleção e sintetização de arquivos de áudio que deram origem ao *dataset* utilizado nos experimentos realizados. O conjunto de dados agrega três tipos de arquivos de áudio com as seguintes características: “eventos”, possui três classes, apenas de sons de alerta; “ambiente”, que possui os mesmos arquivos do primeiro, somados arquivos de som ambiente; e o terceiro “real”, composto por arquivos de som ambiente, sendo que alguns possuem sons de alerta sobrepostos em períodos aleatórios dos arquivos.

Após definições de tipos de sons importantes ao projeto, ocorrida na fase de concepção, foram realizadas buscas em repositórios na *Web* visando coletar arquivos de

[1] *Open Symbols* é uma coleção de símbolos de imagem licenciados abertos, extraídos de várias fontes, que podem ser usados para comunicação aumentativa. Disponível em ⟨https://www.opensymbols.org/⟩.

Figura 29: Símbolos visuais para sugestão de ocorrência de eventos

(a) Ativo, ouvindo



(b) Bebê chorando

(c) Disparo de Arma

(d) Vidro quebrando

Fonte: autor

áudio que atendessem as necessidades de treinamento e testes para os experimentos deste trabalho. A busca não objetivou o maior número de arquivos, mas sim um conjunto que subsidiasse a pesquisa sem acarretar em necessidade de grande estrutura e/ou tempo para realização dos experimentos. Foram selecionados em três momentos distintos, arquivos de datasets que tem origem na Freesound[2] que foram tratados e disponibilizados pela DCASE[3] em evento de detecção e classificação de sons.

No primeiro momento, na busca por arquivos do tipo eventos foram selecionados a partir de desafio de classificação de sons raros. Este desafio abordou diversos tipos de sons (classes), três que coincidiram com os sons importantes aos surdos relatados no levantamento prévio deste projeto, foram selecionadas: choro de bebê, quebra de vidro e disparo de arma de fogo. Numa segunda fase, objetivando arquivos ambiente, um *dataset* disponibilizado a partir de desafio para identificação de cenas acústicas, que abrange arquivos com sons característicos de quinze tipos de lugares, conforme Tabela 4, foi selecionado na íntegra. O terceiro momento coletou um conjunto de arquivos, do tipo áudios da vida real, doravante chamado real, foi identificado e selecionado a partir de desafio de detecção e classificação de eventos acústicos. Este *dataset* possui arquivos de som ambiente que podem conter um som de alerta sobreposto em período aleatório de cada

[2]O Freesound.org é um repositório colaborativo de amostras de áudio com licença CC e organização sem fins lucrativos.

[3]DCASE é o acrônimo para *Detection and Classification of Acoustic Scenes and Events*, promotora o evento de desafios DCASE *Challenge*, oficiais da IEEE (*Institute of Electrical and Electronic Engineers*) e do comitê técnico, pertencente a sociedade de Processamento de Sinais da IEEE, AASP (*Audio e Acoustic Signal Processing*).

arquivo. Por conta do objetivo do projeto, este último *dataset* foi reduzido para aquivos que não apresentam sons de alerta e que apresentam sons correspondentes aos alertas selecionados.

Tabela 4: Composição do *dataset*: sons ambiente

Interior	Ar livre
Biblioteca	Área residencial
Bonde viajando	Centro de uma cidade
Café/Restaurante	Parque urbano
Casa	Praia
Carro em movimento na cidade	Trilha na floresta
Escritório com várias pessoas	
Estação de Metrô	
Mercearia	
Ônibus em movimento na cidade	
Trem em viagem	

Fonte: autor

Os áudios acompanham arquivo contendo anotações sobre seu conteúdo (tipo de ambiente e/ou ocorrência de evento), esta classificação foi realizada, em sua origem, por um anotador humano (MESAROS A.; HEITTOLA, 2017)[4]. Todas as gravações em formato de som WAV, configuradas com uma taxa de amostragem >= que 44100 Hz, foram normalizadas para frequência de 44,1 kHz e resolução de 24-bits.

Mais especificamente, os arquivos de eventos possuem duração variável, que corresponde ao tamanho do evento. Para obter esta configuração, a Freesound realizou uma segmentação semi-supervisionada no *dataset* com um modelo *Support Vector Machine* (SVM)[5] treinado para distinguir entre quadros de alta energia e baixa energia a curto prazo e, em seguida, aplicado em toda a gravação. As estatísticas da duração dos eventos isolados em segundos são as seguintes:

- Choro de bebê, max: 5,1, min: 0,66, média: 2,25s.
- Quebra de vidro, max: 4,54, min: 0,26, média: 1,16s.
- Disparo de arma de fogo, max: 4,4, min: 0,24, média: 1,32s.

Os áudios ambiente são composto por sons de ambientes diversos, foram coletados junto a desafio de classificação de cenas acústicas. Cada arquivo é uma parcela de 10 segundos extraído de capturas de 3 a 5 minutos realizadas pela Freesound. Estes arquivos

[4]O *dataset* foi elaborado e disponibilizado por pesquisadores de laboratório de processamento de sinal da Universidade de Tecnologia de Tampere, na Finlândia, para um desafio de identificação de sons raros.

[5]Uma SVN, máquina de vetores de suporte, é um conceito na ciência da computação para um conjunto de métodos do aprendizado supervisionado que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão.

auxiliam no processo de classificação de eventos enquanto simulam a audição humana em ambientes reais. Mesmo que dividido em quinze classes diferentes (dentro de cafeterias, automóveis e ônibus, restaurantes, ruas movimentadas e parques etc), o motivo da utilização é para que a rede seja capaz de informar que determinado som analisado não corresponde ou contém alguma classe especificada como alerta.

O terceiro conjunto de arquivos, também elaborado pela Freesound, foi adaptado de coleta realizada junto a desafio de detecção de eventos em sons da vida real. Arquivos de áudio do tipo real correspondem a sobreposição de áudios de eventos aos de ambiente. Idealizados para compôr a etapa de testes, simulando situações reais, foi elaborado com cada arquivo contendo 31 segundos de duração de sons ambiente podendo, em alguns casos, conter som de algum evento sobreposto por qualquer trecho do arquivo. Contém um total de 1500 arquivos, divididos em 747 como sons de alerta e 753 de sons ambiente.

A composição do *dataset* que se relaciona a esta pesquisa se dá na junção dos três tipos de arquivos apresentados nesta seção. Abaixo são apresentadas as quantidades de arquivos, tamanhos em segundos e quantidades de classes para cada um deles.

- Eventos: 160 arquivos de tamanho variado, de 3 classes.
- Ambiente: 1874 arquivos de 10 segundos cada, de 15 classes.
- Reais: 1500 arquivos de 31 segundos cada, das 15 classes ambiente, porém, 747 deles possuem o som de uma das 3 classes de evento sobreposta em algum trecho do áudio.

4 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta os experimentos com diferentes configurações de modelos utilizando redes neurais, todos eles treinados e testados utilizando os mesmos arquivos de *datasets*, descritos na seção 3.3. Ao final, são apresentados e discutidos os resultados, comparando desempenho considerando o tipo de arquivo utilizado para classificação.

Apoiado em estudos recentes que apontam arquiteturas de CNNs e RNNs como o estado da arte em AED os experimentos realizados objetivaram, na relação com as soluções apresentadas na seção 2.3.9, soluções envolvendo estas tecnologias. Sendo este um problema de grande complexidade e na ideia de ser aplicada em dispositivos móveis, a resolução por ML poderia, por conta da conectividade total entre as camadas, tornar a rede muito grande e com alto custo de processamento.

Outro motivo observado está na possibilidade de ampliação das classes a serem reconhecidas pela solução. O projeto selecionou três tipos específicos de sons de alerta, mas muitos outros podem ser importantes de serem classificados nos mais variados contextos, o aprendizado profundo pode oferecer maior facilidade no aprendizado para reconhecimento de novos padrões, por exemplo, ao resolverem o problema de ponta a ponta, não necessitando que um especialista realize alterações de extração de recursos específicas aos tipos de padrões que se deseja classificar.

4.0.1 Primeiro Experimento - Conv1 e Conv2

O primeiro experimento utilizou duas redes neurais convolucionais, implementadas em linguagem de programação Python, configuradas a partir de codificação disponibilizada por Zafarullah Mahmood via comunidade Kaggle[1]. Todo este experimento objetiva classificação de eventos sonoros utilizando parte do *dataset* do projeto, neste momento foram utilizados apenas os arquivos de áudio do tipo “evento” e “ambiente”, como descrito na seção 3.3, para as duas redes.

Por conta de redes convolucionais terem tamanho de entrada fixado, tanto em fase de treinamento quanto de testes as redes receberam trechos de 2 segundos (88200 *frames*)

[1] A Kaggle é uma comunidade on-line de cientistas de dados e aprendizes de máquinas, de propriedade da Google LLC. Disponível em ⟨https://www.kaggle.com/⟩.

extraídos aleatoriamente de cada entrada de áudio. Como pode ser verificado no Trecho de Código 4.0.1.1, caso o arquivo de entrada seja menor que o tamanho de segmento fixado, o período faltante é preenchido com dados gerados aleatoriamente.

Trecho de Código 4.0.1.1: Obtendo segmento de áudio e preenchendo com dados aleatórios quando necessário

```
if len(data) > input_length:
    max_offset = len(data) - input_length
    offset = np.random.randint(max_offset)
    data = data[offset:(input_length+offset)]
else:
    if input_length > len(data):
        max_offset = input_length - len(data)
        offset = np.random.randint(max_offset)
    else:
        offset = 0
    data = np.pad(data, (offset, input_length - len(data) - offset),
    "constant")
```

A primeira rede neural, chamada “Conv1” utilizou de camadas de convolução de uma dimensão para a análise das informações dos áudios brutos. Para adequar a informação dos áudios à biblioteca Keras utilizada, em etapa de pré-processamento foi utilizada a função “librosa.load” para gerar matrizes equivalentes aos áudios e seguir para a primeira camada da rede.

A função “librosa.load” realiza a leitura do arquivo de áudio como uma onda e retorna, além da taxa de amostragem, uma série temporal como matriz temporal. Como exemplo, o Trecho de Código 4.0.1.4 mostra informações resultantes da conversão de um segundo dos arquivos de áudio para matriz temporal utilizando a biblioteca librosa. Mais detalhadamente, os áudios foram convertidos para uma frequência de 22050 frames por segundo e ali são apresentadas informações dos três primeiros e três últimos frames desta conversão para arquivos de alerta utilizados no projeto. Nestes áudios, pode ser observado em suas matrizes que o arquivo de choro de bebê tem as informações iniciais como zero, indicando silêncio ao início do áudio, e o vidro quebrando tem logo nas primeiras informações uma amplitude elevada em comparação com o da arma.

Trecho de Código 4.0.1.2: Exemplo de matrizes de áudio

```
#Bebe chorando
[ 0.          0.          0.          ... -0.01726776 -0.01174629
 -0.01236786]

#Vidro quebrando
[ 7.6823622e-02  4.5314482e-01  4.6514377e-01 ...  1.0875671e-05
 -1.8963405e-05  1.5720170e-05]

```

```
#Disparo de arma de fogo
[-0.00027216 -0.0014948  -0.0009736  ... -0.14385936 -0.10651423
  0.15306918]
```

Na estrutura da Conv1 - conforme Apêndice C - , todas as suas camadas convolucionais estão configuradas com função de ativação do tipo ReLU e *padding* considerando apenas informações válidas[2], portanto, são desconsideradas no mapa de recursos resultante aquelas informações que estão nas bordas dos vetores analisados.

Na arquitetura da rede, as camadas iniciais podem ser descritas como três conjuntos contendo duas camadas de convolução de uma dimensão seguidas de uma camada de *max pool* e uma de *dropout* cada. O primeiro conjunto tem 16 filtros de tamanho 9 em cada camada de convolução, saída reduzida ao tamanho 16 na camada de *max pool*[3] e uma função de *dropout*[4] ao final, configurada para eliminar da etapa de treinamento uma fração de 0,1 dos neurônios recebidos na entrada. Tanto o segundo quanto terceiro conjunto de camadas tem 32 filtros em cada camada de convolução, que são de tamanho 3, seguido de função *max pool* de tamanho 4 e mesmo *dropout* da primeira.

Após esta sequência, a Conv1 possui mais duas camadas de convolução de 1 dimensão, cada uma com 256 filtros de tamanho 3 que os resultados dão entrada em uma camada de *max pool* global, que reduzem o filtro ao seu maior valor apresentado. Depois é eliminada a fração de 0,2 dos neurônios em camada *dropout*.

As camadas finais são formadas por duas camadas do tipo *dense* que multiplicam as matrizes de pesos dos neurônios ativos (por meio de ativação ReLU), alterando a dimensão dos vetores para 64 e, na sequência com outra camada *dense*, para 1028. Por fim uma última camada do tipo *dense*, configurada para reduzir a dimensão do vetor de saída para o número de classes treinadas que, por meio da função de ativação *softmax*, apresenta a probabilidade de possíveis resultados para a classificação de cada classe.

A segunda rede, chamada Conv2, tem como uma diferença a utilização de camadas convolucionais de duas dimensões em sua estrutura. Por conta disso, no processo de pré-processamento pra esta rede é realiza a extração do MFCC dos arquivos para posterior envio à primeira camada da rede. Para a extração do MFCC foi utilizada uma função da biblioteca Librosa, escrita em Python, conforme representado no Trecho de Código 4.0.1.3.

Trecho de Código 4.0.1.3: Exemplo de código utilizando biblioteca Librosa para extração

[2]Significa que não adicionamos o preenchimento de informações nulas ao redor da matriz de entrada, apenas o pixel interno da entrada é considerado "válido", que aparece no mapa de recursos.

[3]O método *Max Pool* consiste em reduzir a dimensão das camadas de entrada, pegando o valor máximo de cada região e eliminando valores desprezíveis, cria uma invariância a pequenas mudanças e distorções locais, visando diminuir o custo computacional e evitar *overfitting*.

[4]A função de *Dropout* foi criada para melhorar o treinamento e reduzir o *overffiting*. Na etapa de treinamento da rede esta função exclui certa quantidade de neurônios e os coloca posteriormente. Dessa forma, tem funcionamento parecido com a utilização de mais de uma rede, treinadas individualmente, calculando a média para para a tomada de decisão

do MFCC

```
1 import librosa # utilizacao da biblioteca
2 SAMPLE_RATE = 44100 # frequencia do audio
3 fname = '../input/' + 'audio.wav'   # arquivo de audio de exemplo
4 wav, _ = librosa.core.load(fname, sr=SAMPLE_RATE) # retorno em
      formato de vetor e taxa de amostragem
5 wav = wav[:2*44100] # limitando o vetor para representacao de 2
      segundos de audio
6
7 mfcc = librosa.feature.mfcc(wav, sr = SAMPLE_RATE, n_mfcc=40) #
      gerado MFCC do arquivo
```

Cada arquivo de áudio é passado como argumento para a função MFCC, linha 7 do código 4.0.1.3, juntamente com a configuração de frequência (44100 Hz) e a quantidade de MFCCs, argumento "n_mfcc", que devem ser retornados. Este último argumento representa o tamanho do recurso MFCC, está relacionado à resolução do envelope espectral e influência na riqueza de detalhes. Foram observados trabalhos em que, para o reconhecimento de voz, é usual utilizar tamanho 24 como parâmetro, enquanto para algumas outras aplicações que realmente precisam se aprofundar nos detalhes espectrais é configurado com dimensão 39 ou maior. Como exemplo, o Trecho de Código **??** apresenta as informações de MFCC extraídos de cada tipo de áudio de alerta. Estas são as informações que a rede recebe e analisa. Por conta do tamanho das informações (44100 registros para cada segundo de áudio), aqui são apresentados apenas o primeiro registro para cada um.

Trecho de Código 4.0.1.4: Exemplo de matrizes de áudio

```
1 #Bebe chorando
2 [[-786.60095  -786.60095  -786.60095  -786.60095  -786.60095  -786.60095
3   -786.60095  -786.60095  -786.60095  -786.60095  -614.62415  -499.5273
4   -470.09067  -468.02087  -469.93054  -460.811    -451.0875   -443.5171
5   -453.59967  -465.08453  -472.8262   -490.23776  -503.14026  -512.2283
6   -499.10873  -469.13824  -453.6875   -437.9444   -420.20386  -410.4374
7   -402.48367  -403.7224   -418.42896  -451.90927  -476.4232   -467.68945
8   -454.26413  -446.04834  -432.3068   -427.62524  -408.4074   -391.98938
9   -367.7644   -350.16727]]
10
11 #Vidro quebrando
12 [[   -7.796462   -51.39155   -204.29115   -224.85928   -281.26877
      -329.54993
13   -382.96295   -332.7604    -237.72423   -258.51666   -379.66946
      -450.04758
14   -476.14932   -469.94525   -482.89575   -520.2546    -552.6987
      -562.78674
15   -572.57245   -598.86633   -596.14886   -598.42786   -625.3932
      -629.22614
16   -628.5896    -647.11163   -666.37915   -671.80945   -676.7251
```

```
    -677.1697
17  -674.3266   -656.95935   -661.24994   -675.60864   -675.9921
    -675.8213
18  -673.7083   -669.81604   -674.8632    -678.0038    -679.0361
    -678.71515
19  -678.8061   -679.97144 ]]
20
21 #Disparo de arma de fogo
22 [[-494.97452   -496.9318    -506.69397   -517.0686    -531.4854
    -545.7214
23  -544.23804   -538.2248    -540.70856   -532.89484   -527.6426
    -525.6156
24  -515.464     -510.92557   -517.61127   -517.62604   -525.28925
    -536.41925
25  -536.6381    -537.6814    -536.9383    -539.07007   -541.151
    -542.22876
26  -542.398     -538.94934   -533.6612    -536.2764    -541.8253
    -543.936
27  -542.7047    -541.8563    -536.0985    -534.2024    -529.82794
    -521.56964
28  -513.0001    -509.32474   -507.08386   -513.06616   -131.33173
    16.860819
29   44.795975    26.804386]]
```

A arquitetura da Conv2 é semelhante a Conv1, foi configurada - conforme Apêndice D - de forma que pode ser agrupada em quatro grupos, que contém outras quatro operações cada. Esses grupos possuem uma camada de convolução de duas dimensões com 32 filtros com argumento "*padding = same*", que adiciona informações nulas uniformemente à esquerda e direita do vetor. Os vetores possuem quatro posições com dez posições cada. Depois de cada camada de convolução há uma de *batch normalization* para normalizar a camada de entrada, realizando ajustes no dimensionamento antes de enviar os vetores para a próxima camada, que é de ativação linear, ReLU. A saída desta camada segue para uma camada *max pool*.

Ao final da rede há uma camada *flatten* para a conversão dos vetores de duas dimensões para uma, seguida de uma camada *dense* para diminuir a quantidade de conexões a 64, uma *batch normalization* e outra ReLU de ativação. O resultado desta ReLU segue para outra camada *dense*, que reduz a quantidade de conexões para a quantidade de classes, as quais a rede fará predições e, por fim, chega na última camada, do tipo *softmax*, para a conversão das predições em número indicativo de probabilidade da entrada pertencer a cada classe.

Para estas duas redes o treinamento foi realizado com os arquivos de evento e de ambiente divididos em 75% para treinamento e o restante para teste. A taxa de aprendizado configurada em 0.001, utilizando alguns procedimentos de *callback* para monitorar o trei-

namento, como *ModelCheckpoint*[5], com o *Early Stopping*[6] para monitorar o treinamento, interrompendo o processo quando a rede parar de melhorar seu aprendizado na tentativa de evitar a ocorrência de *underfitting* ou *overfitting*.

A utilização de arquivos de ambiente juntamente com os de alerta se fez necessária, pois não havendo classes que represente ausência de alerta as redes sempre apontarão que algum alerta aconteceu. Dessa forma foram gerados resultados onde podem aparecer nos resultados, por exemplo, predições de falso positivo e falso negativo, possibilitando a utilização das métricas de avaliação de resultados em AED, abordadas na seção 2.3.11.1.

Na combinação dos resultados das predições geradas a partir das redes Conv1 e Conv2 foi gerada um terceiro conjunto de predições. A obtenção deste resultado se deu na multiplicação das predições correspondentes a cada áudio geradas nas redes anteriores. Este novo valor, retornado à proporção dos valores de predição, caracteriza o processo utilizado em combinação de modelos de redes neurais, conhecido como *ensemble averaging*. Após processamento com os arquivos de teste as predições foram salvas em arquivos de modo texto, relacionando a predição (rótulo de classe) para cada arquivo de áudio dado como entrada.

4.0.2 Segundo Experimento - LSTM

Frente ao cenário do projeto, onde os dados que se deseja classificar correspondem a sons de eventos com tamanhos variados, a característica das redes convolucionais de ter suas entradas de tamanho fixado se tornou uma limitação. Como alternativa a este problema, o segundo experimento buscou, assim como em WESTON J. R.; RATLE (2012); SOCHER R.; PERELYGIN (2013); SOCHER R.; HUANG (2011), a utilização de uma rede com estrutura recorrente, encontrando uma rede LSTM baseada na rede MobileNetV2 disponibilizada na comunidade Kaggle por Sainath Adapa.

A utilização de uma *Mobile Network* (MobileNet) também foi desejado por esta rede ter como pilar a eficiência para consumir menos recursos computacionais, sua característica principal é ser apropriada para aplicações em smartphones e embarcadas. A MobileNet utilizada neste experimento é a versão 2, que obteve 8º lugar (no Kaggle) para o desafio DCASE 2018 de distinguir 41 tipos diferentes de sons usando arquivos WAV fornecidos que incluem coisas como instrumentos musicais, sons humanos, sons domésticos e animais.

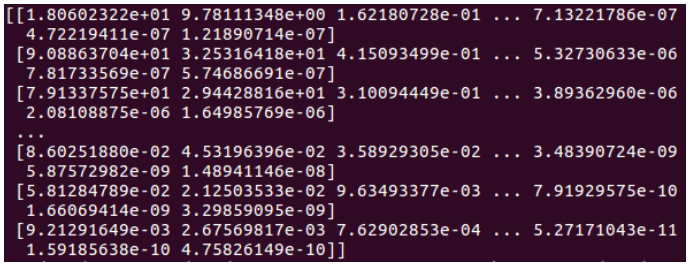
Esta solução é projetada para remover o silêncio nas extremidades dos áudios, realiza esta operação na função "trim" da biblioteca Librosa, extrair o MFCC dos arquivos e os dados do *Log Mel-Spec* de entrada, função "librosa.feature.melspectrogram", que tem como resultado uma matriz como mostra a Figura 30, e, ainda antes de enviar os dados

[5]O *ModelCheckpoint* salva o melhor peso do modelo (usando dados de validação) para fazer previsões de teste.

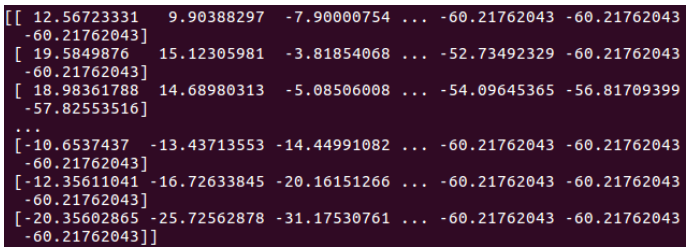
[6]O *Early Stopping* ou "parada antecipada" é um tipo de método de regularização utilizado em redes neurais na tentativa de impedir o *overfitting*.

para a MobileNetV2, converte o espectrograma de potência (amplitude ao quadrado) em unidades de decibéis (dB) pela função “power_to_db”, também da Librosa. Esta última função, em especial, realiza a conversão calculando a escala de maneira numericamente estável através da fórmula 10 * log10(S / ref), onde S é uma potência de entrada e ref a amplitude.

Figura 30: Representação em trecho de matrizes de áudio de vidro quebrando



(a) Matriz do MFCC



(b) Matriz do MFCC em decibéis (db)

Fonte: autor

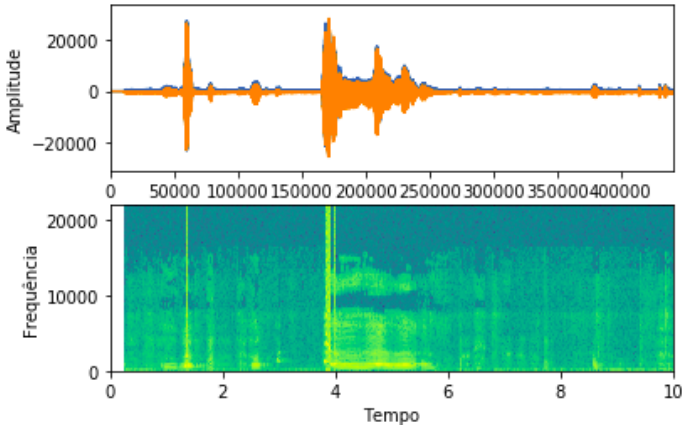
Após este processamento os dados são enviados a MobileNetV2 e sua saída enviada para uma rede convolucional utilizada para realizar a redução de dimensionalidade da rede por meio de uma série de camadas densas. Para a realização deste experimento foram realizados ajustes na rede quanto ao propósito da classificação, como quanto à dimensão de classes. Foi realizado treinamento utilizando apenas os arquivos de alerta do *dataset* deste estudo e foram realizados dois testes, um com os arquivos de evento e outro com os arquivos de som real. As predições foram salvas em arquivos de modo texto, relacionando a predição (rótulo de classe) para cada arquivo de áudio dado como entrada.

4.0.3 Análise e Resultados

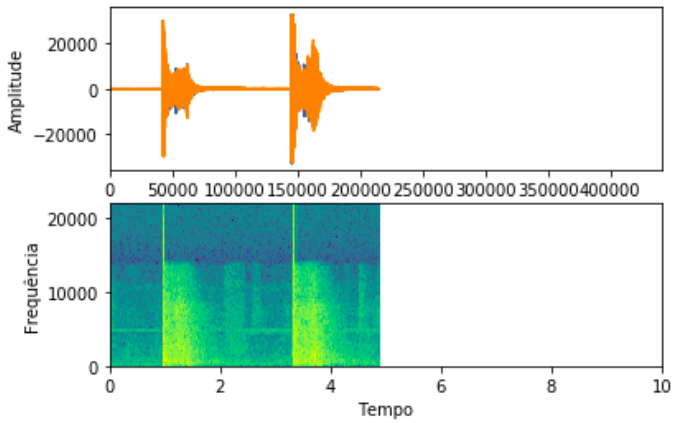
Este capítulo apresenta os resultados obtidos a cerca dos experimentos realizados que envolveram técnicas de extração de características em arquivos de áudio e o processamento em redes neurais, assim como, relacionando os resultados com as características principais das estruturas e comparando seus resultados. A Figura 31 mostra representações de um arquivo de áudio de cada tipo de som de alerta. Esta representação será referenciada no decorrer desta seção como forma de auxiliar na análise e discussão

de resultados.

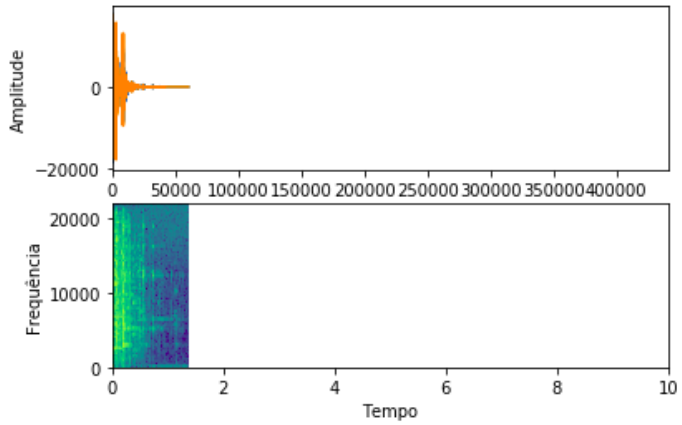
Figura 31: Representação gráfica de áudio em forma de onda (Amplitude x Tempo[amostra]) e espectrograma (Frequência[kHz] x Tempo[s])



(a) Choro de bebê



(b) Tiro de arma de fogo



(c) Vidro quebrando

Fonte: autor

O primeiro experimento obteve três conjuntos de resultados, um de cada rede testada (Conv1 e Conv2) e um terceiro, chamado “Conv1e2”, gerado na combinação dos ante-

riores. A Tabela 5 apresenta matrizes de confusão, onde, na relação de tipo de arquivo com a predição obtida e considerando os retornos de alerta como positivos, são expressos o percentual de verdadeiro positivos, falsos positivos, verdadeiros negativos e falsos negativos.

Tabela 5: Primeiro experimento: matriz de confusão

(a)

Conv1			
		Classe Verdadeira	
		Alerta	Ambiente
Predição	Alerta	0	0
	Ambiente	100%	100%

(b)

Conv2			
		Classe Verdadeira	
		Alerta	Ambiente
Predição	Alerta	87,8%	0,1%
	Ambiente	12,19%	99,9%

(c)

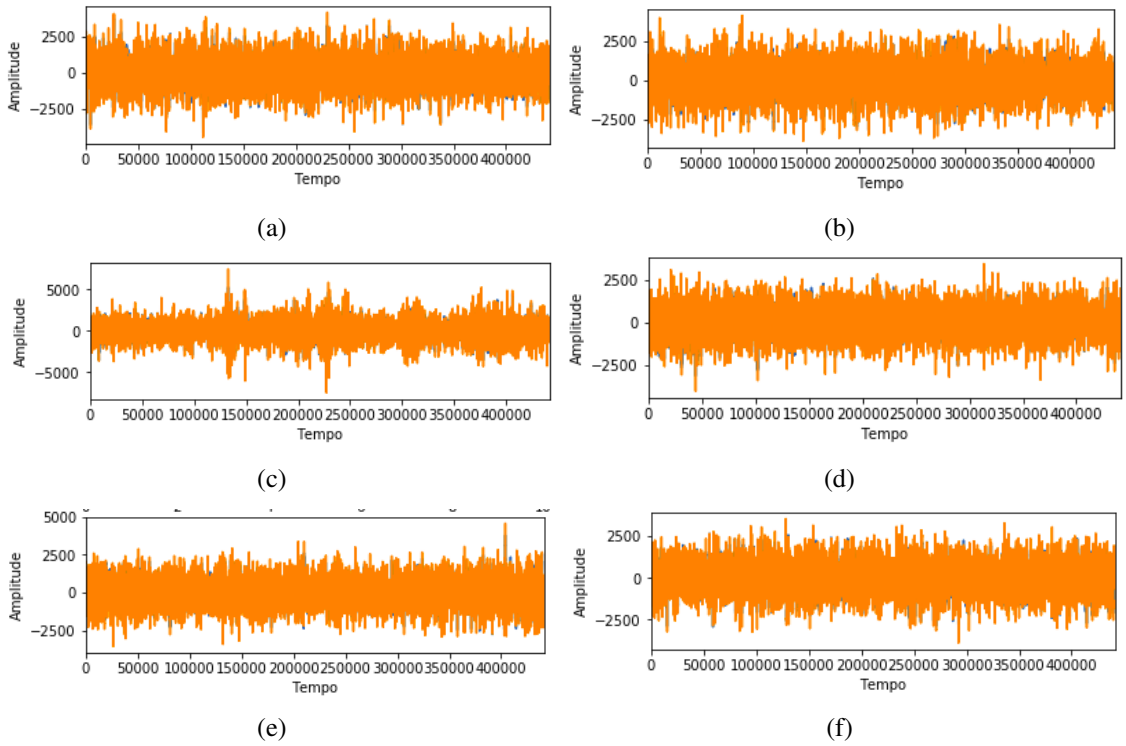
Conv1e2			
		Classe Verdadeira	
		Alerta	Ambiente
Predição	Alerta	80,48%	0,06%
	Ambiente	19,51%	99,94%

Fonte: autor

As predições da rede Conv1 apontaram som classificado ambiente neste estudo para todos os arquivos testados, mais especificamente, para a classe de som de trem. Dentre todos os testes realizados, esta foi a única configuração de rede que não utilizou informações extraídas de MFCC para classificação. Observando os gráficos que apresentam a amplitude do sinal no tempo dos áudios, Figura 32, não foi identificado padrão de onda semelhante às características dos áudios de alerta (Figura 31). A audição de dez dos arquivos de trem em viagem do *dataset* apontaram, como pode ser verificado nos gráficos, ruídos constantes e eventuais eventos, como fala e eventos, aparentemente, de choque em madeira. Com base nisso, as predições realizadas pela Conv1 sugerem que os trechos de áudio dos arquivos de ambiente das demais classes utilizados no treinamento não apresentavam altas amplitudes (Apêndice B), contrário das classes de eventos que em trechos de um segundo, como utilizado nesta rede, basicamente, não apresentam trechos sem altas amplitudes.

Os resultados da rede Conv2 apontaram uma melhora nas predições para os dois tipos de som (alerta e ambiente), alcançando índice de acerto de 99,9% para sons de ambiente e 87,8% para alertas. A predição gerada a partir da combinação das anteriores (Conv1e2) apresentou índices percentuais aproximados que da Conv2, as duas utilizaram características de espectrogramas para treinamento e predições. É possível observar que, em comparação às características analisadas pela Conv1, a configuração visual de espectrograma dos áudios de trem, como no exemplo da Figura 33 que representa os arquivos correspondentes aos da Figura 32, as informações de espectrograma atenuam as características dos ruídos em relação a sua amplitude. Desta forma, os espectrogramas de trem possuem aparência mais uniforme, como visto em áudios de ambiente (Apêndice B), e não tanto como os de eventos, como quando analisado apenas suas amplitudes. Insto indica que a utilização de informações de MFCC proporcionaram melhores resultados relação a apenas informações sobre a amplitude do sinal.

Figura 32: Gráficos de áudio em forma de onda de trem em viagem (Amplitude x Tempo[amostra])



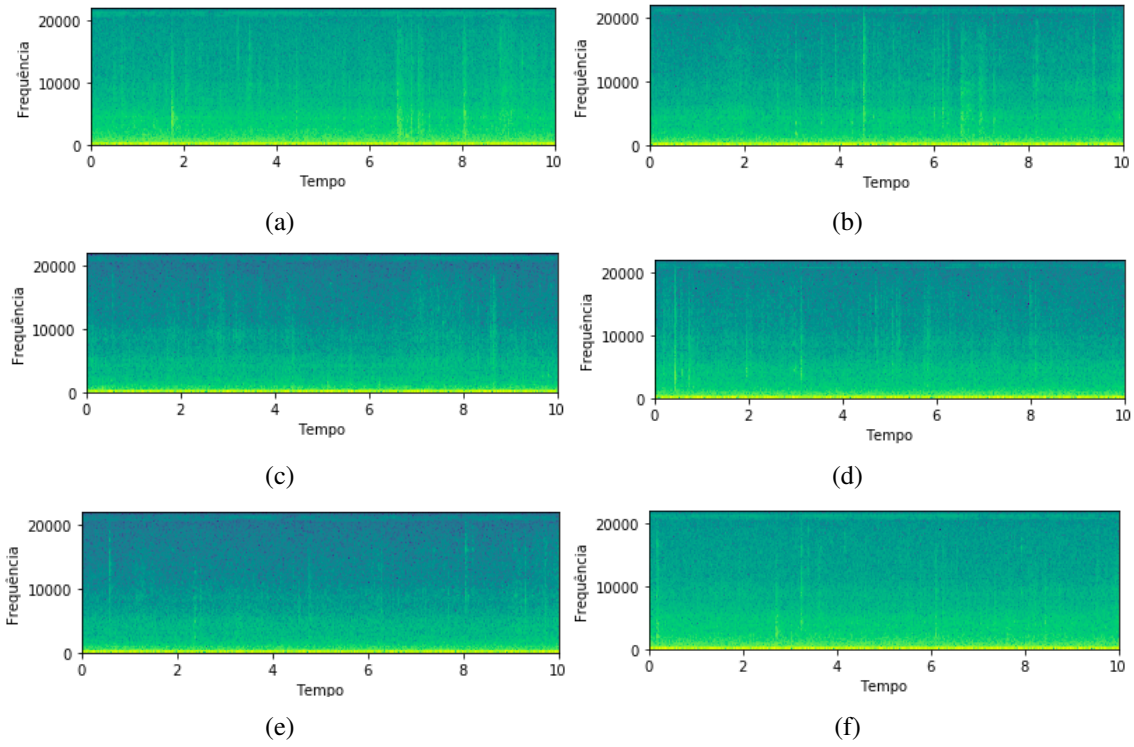
Fonte: autor

O segundo experimento, que utilizou uma rede LSTM, obteve dois conjuntos de predições, uma a partir de testes com conjunto de dados de eventos e outro com arquivos de som real, estes testes são referenciados, respectivamente, como LSTM e LSTM real. O primeiro teste realizado com a rede LSTM, treinada com sons de alerta e ambiente, utilizou apenas arquivos de alerta para as predições e obteve acerto de 86,66% das classes de eventos dos arquivos, este índice aponta a aprendizagem em relação aos alertas, mas impossibilita a geração de índices como o *F1 Score*, pois o tipo de teste não possibilita a geração de falsos positivos e falsos negativos.

O teste utilizando arquivos reais (sons ambiente podendo haver ocorrência de alerta) na mesma rede LSTM treinada gerou a matriz de confusão expressa na Tabela 6. Neste teste a rede apontou resultados inferiores aos da Conv2, por exemplo, porém o tipo de teste é de natureza diferente. Enquanto a rede Conv2 necessita que seja enviado um trecho de áudio de tamanho especificado, a rede LSTM recebe os arquivos com tamanho variável, situação mais aproximada de uma captação de áudio na vida real. Este teste apontou acertos de apenas 28,64% para existência de alerta e 36,65% para arquivos de ambiente.

Como pode ser visto na Figura 34, visualmente os espectrogramas de som ambiente que contém ocorrência dos alertas definidos no projeto apresentam características notáveis até mesmo para a observação humana. Nesta imagem que representa um exemplo de 30

Figura 33: Espectrogramas de áudio de trem em viagem (Frequência[kHz] x Tempo[s])



Fonte: autor

Tabela 6: Segundo experimento: matriz de confusão

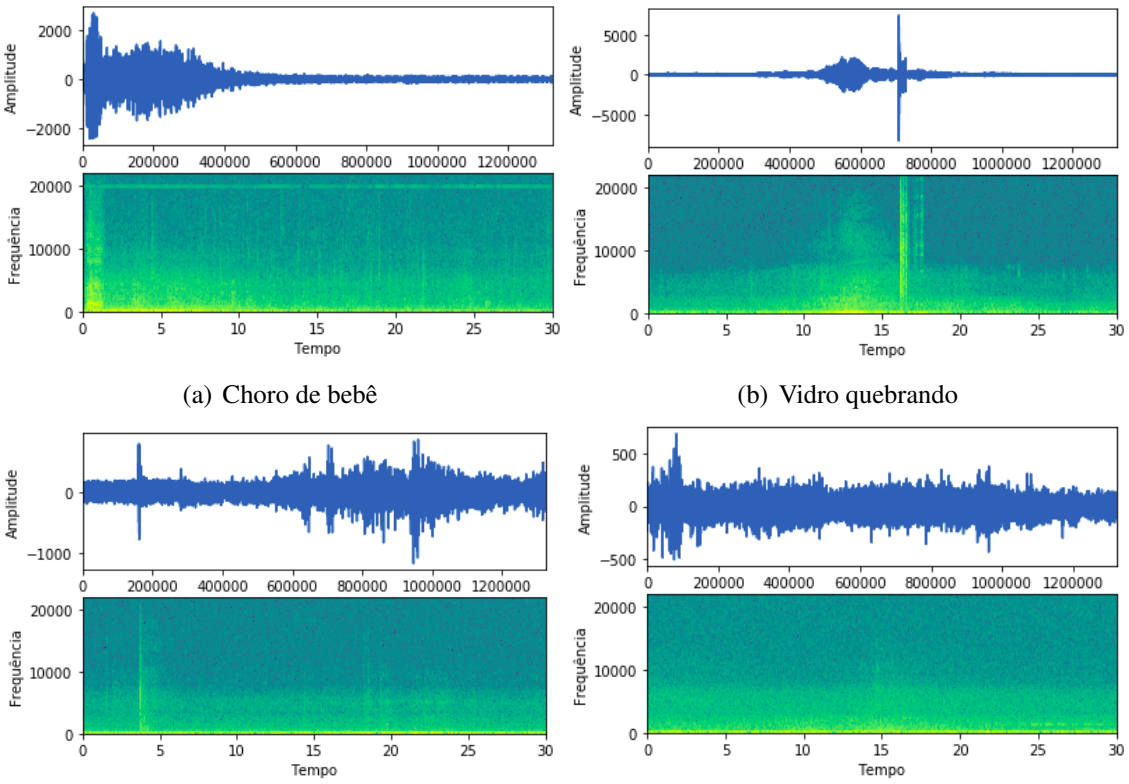
LSTM com arquivos reais			
		Classe Verdadeira	
		Alerta	Ambiente
Predição	Alerta	28,64%	63,35%
	Ambiente	71,26%	36,65%

Fonte: autor

segundos para cada classe, é possível observar, assim como na audição dos arquivos, a ocorrência de evento no início do gráfico que indica choro de bebê, por volta do segundo 16 no arquivo de vidro quebrando, disparo de arma de fogo anterior ao segundo 5 do arquivo e uma imagem sem linhas verticais de cores mais quentes que poderiam indicar ocorrência de algum evento.

Dessa forma, a análise visual dos espectrogramas não sugere a razão para os resultados de acertos abaixo dos 40% para esta rede. Contudo, o percentual de erro desses resultados pode ter ocorrido devido a quantidade de arquivos de som ambiente frente aos de alerta. Na etapa de treinamento foram utilizados 907 arquivos de alerta e 2627 de som ambiente. Para testes e predições foram utilizados 747 arquivos de alerta e 749 de ambiente. A diferença entre a quantidade de arquivos utilizados para treinamento que contém evento de alerta (907) e de ambiente (2627) é de 2,89, esta proporção tem semelhança aos

Figura 34: Espectrogramas de áudio da vida real (Frequência[kHz] x Tempo[s])



(a) Choro de bebê

(b) Vidro quebrando

(c) Disparo de arma de fogo

(d) Sem evento de alerta

Fonte: autor

das predições na razão de erros pelos acertos de cada classe, que é de aproximadamente 2,12 para predições de alerta e 1,94 para de ambiente. Estas informações sugerem que a disparidade das quantidades de arquivos para cada classe na etapa de treinamento pode ter influenciado negativamente nos resultados.

A Tabela 7 apresenta o percentual de acertos de cada teste realizado em relação aos tipos de alertas objetivados. Nela, é possível identificar que, no geral, os arquivos contendo sons de vidro quebrando foram melhores identificados que os demais e em quase todos o som de arma foi melhor identificado que o de choro de bebê. Entre todos os testes, a LSTM obteve o maior índice de acerto para classificação de vidro quebrado e, desconsiderando a Conv1 que não identificou alerta algum, o menor índice no geral, sendo para para arma de fogo.

Analisando os resultados gerais em síntese com as características percebidas visualmente em amostras de representações gráficas dos áudios utilizados é possível sugerir que os sons de vidro quebrando foram melhor identificados em todos os testes, pois apresenta grande amplitude de sinal logo no início da ocorrência do evento. Isto pode ter sido determinante nas redes que receberam os dois primeiros segundos de áudio (Conv1 e Conv2). Já no teste com a LSTM os outros tipos de alerta também obtiveram índices acima de

Tabela 7: Predições corretas de experimento por tipo de alerta

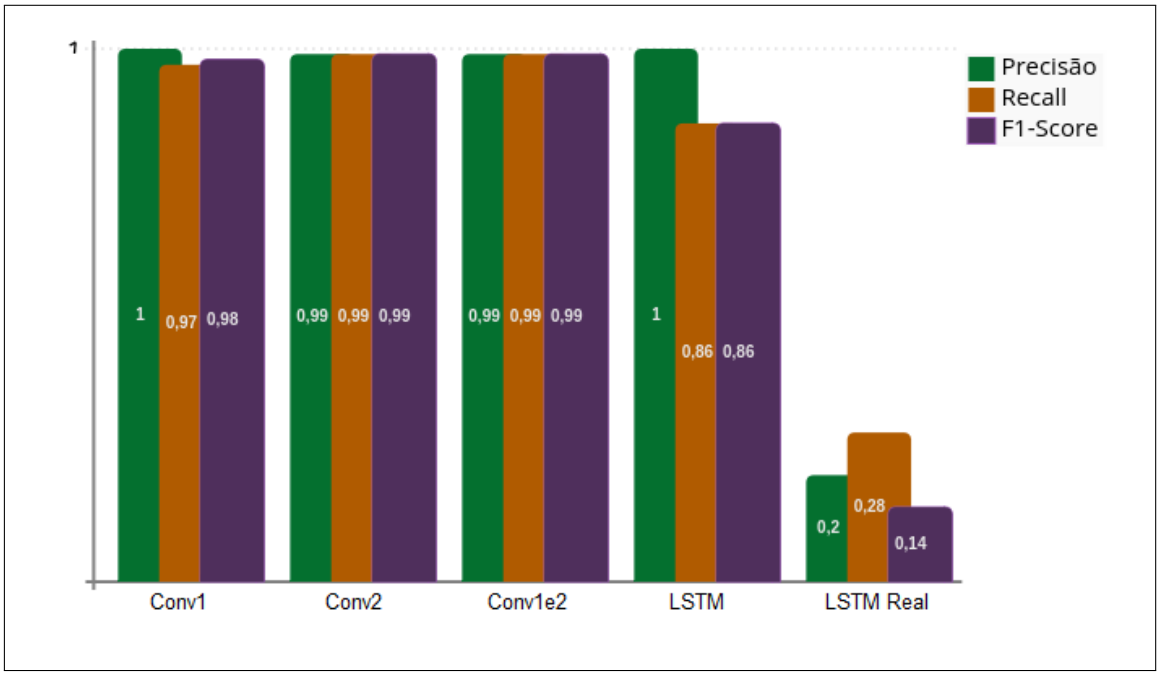
Tipo de alerta \Experimento	Conv1	Conv2	Cpnv1e2	LSTM	LSTM Real
Arma de fogo (%)	0	86,66	82,53	82,6	25,2
Choro de bebê (%)	0	80	78,75	76	25,91
Vidro quebrado (%)	0	93,75	93,75	95,91	34,8

Fonte: autor

76% e na LSTM Real os acertos foram mais baixos para as três classes, não apontando evidência de relação ao tipo de espectrograma para estas redes.

Com os valores gerais obtidos, analisados nesta seção, como nos expostos nas Tabelas 5 e 6, foram calculados, com base no exposto na seção 2.3.11.1, precisão, *recall* e *F1 Score* para cada arquitetura experimentada. Esses cálculos, representados na Figura 35, apresentam altos índices para os quatro primeiros testes, porém, estes não configuraram situações semelhantes às de situações reais. O teste que mais se aproxima de situação real obteve *F1 Score* inferior ao de soluções que, atualmente, representam o estado da arte. Estes índices se relacionam com os percentuais de acertos, analisados no decorrer desta seção, e os apontamentos realizados aos resultados se aplicam da mesma forma aos coeficientes.

Figura 35: Precisão, *recall* e *F1 Score* dos testes



Fonte: autor

5 CONSIDERAÇÕES

Este capítulo apresenta considerações referentes a experimentos realizados na utilização de redes neurais para a detecção e classificação de sons de alerta para pessoas surdas. Abordando assuntos como ensino de pessoas surdas, linguagem visual para a sugestão de formas de comunicação e detecção de eventos sonoros utilizando redes neurais.

A consulta apresentada na seção de Levantamentos Prévios (3.1) apontou que uma tecnologia como da natureza deste trabalho pode ser importante para o dia a dia da pessoa surda. A utilização em sala de aula de tecnologias que auxiliem as pessoas no processo natural de atenção seletiva podem contribuir para o processo de aprendizagem. E a forma de comunicação comunicação visual desta tecnologia proposta, sistema de alertas, deve utilizar elementos com alto fator de pregnância, preferencialmente com símbolos socialmente conhecidos.

As cores também podem contribuir para a transmissão de mensagens, antes mesmo da identificação de outros elementos apresentados. Dessa forma, a cor vermelha simbolizando - dentre outras coisas - alerta, esta deve ser predominantemente utilizada nos símbolos utilizados para sugerir a ocorrência de situações de perigo. Com base nisso, a utilização da cor predominantemente vermelha em sua saturação máxima representa uma das características indicadas a serem observadas no desenvolvimento de uma solução. Desta forma, quatro símbolos foram coletados em repositório disponível na *Internet* e configurados como sugestão de comunicação.

O primeiro experimento, quando a diferença entre dois testes ocorreu em etapa de pré-processamento, apontou que, assim como na literatura, a etapa de pré-processamento pode melhorar significativamente os resultados de classificação da rede neural. A utilização de arquivos brutos no formato wav em uma rede convolucional de uma dimensão (chamada Conv1) não foi capaz de identificar quaisquer evento de alerta do *dataset*. Em comparação, a estrutura da rede Conv2, que se difere da anterior em extrair os MFCCs dos arquivos e submetê-los às camadas convolucionais de duas dimensões da rede obteve acerto de predições para além de 86%. Vale destacar que tal resultado, obtido a partir de uma gama de menos de 200 arquivos de *dataset*, pode caracterizar *overftting*, mais

estudos são necessários para subsidiar este apontamento.

Como no primeiro experimento as redes receberam um tamanho fixo, parte do áudio original, os testes do segundo experimento utilizaram uma rede LSTM que, no primeiro teste pode ser comparado com as redes convolucionais, obteve percentuais de predições corretas não distante de quatro percentuais nos resultados. Esta variância atribui melhor desempenho à rede LSTM em comparação às outras, pois utiliza de tamanho variável de arquivo de entrada, aproximando a solução de uma situação real. Já o segundo teste com a rede LSTM, utilizando áudios reais, obteve desempenho abaixo em comparação ao encontrado na literatura e sugere ocorrência do problema de *underfitting*.

Porém, é importante analisar que para a obtenção destes resultados não foram simuladas situações reais, e sim por arquivos de áudio contendo gravações. Portanto, apesar dos resultados apontarem capacidade na identificação de sons de alertas importantes às pessoas surdas, objetivo principal deste trabalho, para que o modelo viabilize uma aplicação real ainda se faz necessário maior investigação e estudo dos procedimentos que envolveram o processo de detecção e classificação de eventos acústicos.

5.1 Trabalhos Futuros

Os resultados alcançados nesta pesquisa são considerados como ponto de partida para trabalhos futuros que podem ser desenvolvidos visando, tanto melhorias nas soluções desenvolvidas, quanto novas funcionalidades que não foram abordadas neste trabalho. Esta seção apresenta algumas sugestões de estudos que podem colaborar para o desenvolvimento de soluções para outros aspectos do projeto, colaborar para o amadurecimento e alcançar o objetivo motivador desta pesquisa, a constituição uma aplicação que auxilie pessoas surdas.

Com base nos estudos realizados, dificuldades encontradas, implementações e resultados, são sugestivos como trabalhos futuros, por exemplo, estudo buscando verificar se os testes do primeiro experimento apresentaram o problema de *overfitting* e se a rede LSTM, que apresentou capacidade para identificação de eventos em áudios que reproduzem situações reais, e o *dataset* podem ser aprimorados e, também, testado com os áudios em resoluções diferentes visando melhorar os resultados.

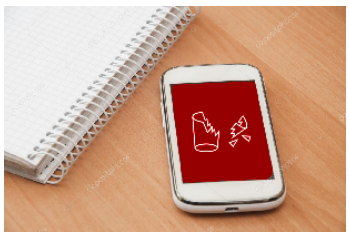
A investigação das razões para que a rede LSTM apresente diferente desempenho quando testada apenas com arquivos de evento em comparação quando testada com arquivos reais. É sugerida a adição de um mecanismo seletor que realize a identificação de ocorrência de evento, em que trecho de áudio está ocorrendo um evento, e envie apenas as informações do evento para outra rede, como a LSTM aqui utilizada, realizar a classificação.

Outra sugestão é a realização de estudos e para a aplicação em *mobile*. Podendo esta, tomar proveito de sensores disponíveis em *smartphones* (*phone sensing*), como o

microfone. Para a comunicação é sugerida a avaliação junto a comunidade surda dos símbolos sugeridos e elaboração de comportamento de um sistema sendo executado em celular, conforme sugere a a Figura 36.

Figura 36: Representação de dispositivo móvel com aplicação ativa

(a) Nenhum alerta identificado



(b) Sugestão de vidro quebrando

Fonte: autor

Outras ideias de trabalhos, sugeridos em banca, podem ser futuramente desenvolvidos com base neste projeto, como na alteração de contexto, generalizar o contexto de aplicação da tecnologia ou especializar mais o ambiente, por exemplo, focar no educacional, buscando ”sons educacionais”. Na alteração de escopo de projeto, considerar alerta por vibração e outros tipos. E no planejamento de uma solução, envolvendo escola bilíngue para os levantamentos prévios.

REFERÊNCIAS

ALEXANDRE D. S.; TAVARES, J. M. R. S. Factores da Percepção Visual Humana na Visualização de Dados. **CMNE 2007 - Congresso de Métodos Numéricos em Engenharia, XXVIII CILAMCE - Congresso Ibero Latino-Americano sobre Métodos Computacionais em Engenharia**, [S.l.], 2007.

AUSTREGESILO, L. E. L. **Desculpe, não ouvi!** São Paulo: Atitude Terra, 2014.

BENGIO, T. **Learning Deep Architectures for AI**. Dept. IRO, Université de Montréal. C.P. 6128, Montreal, Qc. Canadá: now Publishers Inc., 2009.

BENGIO Y.; BOULANGER-LEWANDOWSKI, N. P. R. Advances in optimizing recurrent networks. **Proc. IEEE Int. Conf. Acoust. Speech Signal Process**, [S.l.], p.8624–8628, 2013.

BENGIO Y.; LECUN, S. Learning Algorithms Towards AI. **MIT Press**, [S.l.], 2007.

BISHOP, C. M. **Pattern recognition and machine learning**. [S.l.]: springer, 2006.

BRASIL. **Lei 10.436. Dispõe sobre a Língua Brasileira de Sinais - Libras e dá outras providências**. Disponível em:⟨http://www.planalto.gov.br/ccivil_03/leis/2002/l10436.htm⟩.

BRASIL. **Decreto 5.626. Regulamenta a Lei no 10.436, de 24 de abril de 2002, que dispõe sobre a Língua Brasileira de Sinais - Libras, e o art. 18 da Lei no 10.098, de 19 de dezembro de 2000**. Disponível em:⟨http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5626.htm⟩.

BRASIL. **Lei 13.409. Altera a Lei nº 12.711, de 29 de agosto de 2012, para dispor sobre a reserva de vagas para pessoas com deficiência nos cursos técnico de nível médio e superior das instituições federais de ensino**. Disponível em:⟨http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2016/Lei/L13409.htm⟩.

CAKIR E.; HEITTOLA, T. H. H. V. T. Polyphonic sound event detection using multi label deep neural networks. **IEEE International Joint Conference on Neural Networks (IJCNN)**, [S.l.], 2015.

CANTO, D. S. D. DETERMINACIÓN AUTOMÁTICA DE MODOS MULTI-ARMÓNICOS EN ESPECTROGRAMAS. , E.T.S.I.I Universidad Nacional de Educación a Distancia., 2012.

CARVALHO, D. d. V. **Breve História dos Surdos no Mundo**. [S.l.]: SurdUniverso, 2007.

DAROQUE S. C.; PADILHA, A. M. L. **Alunos Surdos no Ensino Superior**: uma discussão necessária. Universidade Metodista de Piracicaba, Piracicaba-SP: Dissertação (mestrado em educação). Programa de pós-graduação em educação, 2011.

DENG, L. A tutorial survey of architectures, algorithms, and applications for deep learning. **APSIPA Trans. Signal Inf. Process.**, [S.l.], v.3, n.2, p.1–29, 2014.

DENNIS J.; TRAN, H. D. C. E. S. Overlapping sound event recognition using local spectrogram features and the generalised hough transform. **Pattern Recognition Letters**, [S.l.], v.34, n.9, p.1085–1093, 2013.

DONDIS, D. **A Sintaxe da Linguagem Visual**. São Paulo: Martins Fontes, 1991.

ESPI M.; FUJIMOTO, M. K. K. N. T. Exploiting spectro-temporal locality in deep learning basedacoustic event detection. **EURASIP J. Audio Speech Music Process.**, [S.l.], 2015.

FADLULLAH Z. M.; TANG, F. M. B. K. N. A. O. I. T. M. K. State-of-the-art deep learning: evolving machine intelli-gence toward tomorrow's intelligent network traffic control systems. **IEEE Communications Surveys Tutorials**, [S.l.], v.19, n.4, p.2432–2455, 2017.

GERS, F. A.; SCHMIDHUBER, J.; CUMMINS, F. Learning to forget: continual prediction with lstm. , [S.l.], 1999.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016.

GRANDJEAN, E. **Manual de ergonomia**: adaptando o trabalho ao homem. [S.l.]: Bookman, 1998.

HASAN M. R.; JAMIL, M. R. M. G. R. M. S. Speaker indentification using MEL frequency cepstral coefficients. **3rd Int. Conf. on Electrical Computer Engineering ICECE**, [S.l.], 2004.

HEITTOLA T.; MESAROS, A. V. T. G. M. Supervised model training for overlapping sound events based on unsupervised source separation. **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, [S.l.], p.8677–8681, 2013.

HOCHREITER, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. **Int. J. Uncertainty Fuzziness Knowl. Based Syst.**, [S.l.], p.107–116, 1998.

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

HUSSAIN, F.; HASSAN, S. A.; HUSSAIN, R.; HOSSAIN, E. Machine Learning for Resource Management in Cellular and IoT Networks: potentials, current solutions, and open challenges. **arXiv preprint arXiv:1907.08965**, [S.l.], 2019.

IBGE, I. B. d. G. e. E. **Censo demográfico** : 2010 : características gerais da população, religião e pessoas com deficiência. Rio de Janeiro: IBGE, 2010. Disponível em: ⟨https://biblioteca.ibge.gov.br/visualizacao/periodicos/94/cd_2010_religiao_deficiencia.pdf⟩. Acesso em: out. 2017.

JONES, S. **Alerting devices**. "Disponível em:⟨https://www.healthyhearing.com/help/assistive-listening-devices/alerting-devices⟩.".

LADEWIG, I. A importância da atenção na aprendizagem de habilidades motoras. **Revista Paulista de Educação Física**, [S.l.], v.3, p.62–71, 2000.

LIMA, M. C. M. P. et al. **Avaliação de fala de lactentes no período pré-linguístico**: uma proposta para triagem de problemas auditivos. [S.l.]: Tese (Doutorado em Educação) Universidade Estadual de Campinas – UNICAMP, 1997.

LOPES, M. C. **Surdez Educação**. Belo Horizonte: Autêntica, 2007.

MARRONI, L. S. **Aplicação da Transformada de Hough para localização dos olhos em faces humanas**. 2002. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo.

MASEK, L. et al. **Recognition of human iris patterns for biometric identification**. 2003. Tese (Doutorado em Ciência da Computação) — Master's thesis, University of Western Australia.

MEC/INEP/DEEP. **Censo da Educação Superior 2017**: divulgação dos principais resultados. Disponível em:⟨http://portal.mec.gov.br/docman/setembro-2018-pdf/97041-apresentac-a-o-censo-superior-u-ltimo/file⟩.

MESAROS, A. et al. DCASE 2017 challenge setup: tasks, datasets and baseline system. **Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)**, [S.l.], nov 2017.

MESAROS A.; HEITTOLA, T. D. A. E. B. S. A. V. E. R. B. V. T. Dcase 2017 challenge setup: tasks, datasets and baseline system. **Proceedings of DCASE2017 Workshop**, [S.l.], 2017.

MESAROS A.; HEITTOLA, T. E. A. V. T. Acoustic event detection in real life recordings. **18th European Signal Processing Conference**, [S.l.], p.1267–1271, 2010.

MESAROS A.; HEITTOLA, T. V. T. Metrics for Polyphonic Sound Event Detection. **Applied Sciences**, [S.l.], v.6, p.162, 2016.

MIKOLOV T.; KARAFIáT, M. B. L. C. J. K. S. Recurrent neural network based language model. **Proc. INTERSPEECH**, [S.l.], p.1045–1048, 2010.

MOURA, D. R. Introdução à Surdez e à Libras. **Enap Escola Nacional de Administração Pública - Diretoria de Comunicação e Pesquisa**, [S.l.], 2016.

MUDA, L. B. M. E. I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. **Journal Of Computing**, [S.l.], v.2, p.138–143, 2010.

ORDONEZ F. J.; ROGGEN, D. Deep convolutional and lstm recurrentneural networks for multimodal wearable activity recog-nition. **Sensors**, [S.l.], 2016.

PANG Y.; SUN, M. J. X. L. X. Convolution in convolution for network in network. **IEEE Transactions on Neural Networks and Learning Systems**, [S.l.], p.1–11, 2017.

PARASCANDOLO G.; HUTTUNEN, H. V. T. Recurrentneural networks for polyphonic sound event detection in reallife recordings. **2016 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP)**, [S.l.], p.6440–6444, 2016.

PIZZANI L.; ROSEMARY, C. S. B. S. F. A arte da pesquisa bibliográfica na busca do conhecimento. **Revista Digital de Bibliotecnia e Ciência da Informação**, [S.l.], v.10, n.1, p.53–66, 2012.

RAZAK Z.; IBRAHIM, N. J. T. E. M. I. M. Y. I. Quarnic Verse recitation feature extraction using Mel-Frequency Cepstral Coefficient(MFCC). **Department of Al-Quran Al-Hadith. Academy Of Islamic Studies, University of Malaya**, [S.l.], 2008.

RENSINK, R. A. **Internal vs. external information in visual perception**. ACM International Conference Proceeding Series: Proceedings of the 2nd international symposium on Smart graphics, 2002. 63-70p. v.24.

RIFAI S.; BENGIO, Y. V. P. D. Y. N. A generative process for sampling contractive auto-encoders. **Proc. 29th Int. Conf. Mach. Learn. (ICML)**, [S.l.], p.1855–1862, 2012.

SEIDE, F. **MAVIS - Microsoft Research**. ”Disponível em:⟨https://www.microsoft.com/en-us/research/project/mavis/⟩.”.

SEIDE F.; LI, G. C. X. Y. D. Feature engineering in context-dependent deep neural networks for conversational speech transcription. **IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)**, [S.l.], p.24–29, 2011.

SHALEV-SHWARTZ S.; BEN-DAVID, S. **Understanding MachineLearning**: from theory to algorithms: from theory to algorithms. [S.l.]: Cambridge University Press, 2014.

SILMAN, S. Auditory diagnosis. **Applied Sciences**, [S.l.], p.49–51, 1991.

SKINNER, B. F. **Tecnologia do Ensino**. São Paulo: Herder; USP, 1972.

SOCHER R.; HUANG, E. H. P. J. N. A. Y. M. C. D. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. **Advances in Neural Information Processing Systems**, [S.l.], p.801–809, 2011.

SOCHER R.; PERELYGIN, A. W. J. Y. C. J. M. C. D. N. A. Y. P. C. Recursive deep models for semantic compositionality over a sentiment treebank. **Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)**, [S.l.], v.1631, p.1–12, 2013.

SUTSKEVER, I. **Training recurrent neural networks**. Univ. Toronto, Toronto, ON, Canada: Ph.D. dissertation, Dept. Comput. Sci., 2013.

TERRA, C. L. **O processo de constituição das identidades surdas em uma escola especial para surdos sob a ótica das três ecologias**. Universidade Federal do Rio Grande, Rio Grande-RS: 188 f. Dissertação (mestrado em educação ambiental). Programa de pós-graduação em educação ambiental, 2011.

THOMÉ, A. C. G. Redes neurais: uma ferramenta para kdd e data mining. **NCE. UFRJ.**, [S.l.], 2002. Disponível em: ⟨http://equipe.nce.ufrs.br/thome/grad/nn/mat_didatico/apostila_kdd_mbi.pdf⟩. Acesso em: jul. 2019.

UTGOFF P. E.; STRACUZZI, D. J. Many-layered learning. **Neural Comput**, [S.l.], v.14, n.10, p.2497–2529, 2002.

VILELA M.; KOCH, I. V. **Gramática da língua portuguesa**. Coimbra: Almedina, 2001.

VIRTANEN, A. M. T. H. T. TUT database for acoustic scene classification and sound event detection. **24th Acoustic Scene Classification Workshop 2016 European Signal Processing Conference (EUSIPCO)**, [S.l.], 2016.

W3C, W. W. W. C. **Módulo de cor CSS Nível 3**. "Disponível em:⟨https://www.w3.org/TR/css-color-3/#hsl-color⟩.".

WARE, C. **Information Visualization**: perception for design. San Francisco: Morgan Kaufmann Publisher, 2004.

WDF, W. F. o. t. D. **WFD Policy**: education rights for deaf children. Disponível em:⟨http://www.wfdeaf.org/pdf/policy_child_ed.pdf⟩.

WESTON J. R.; RATLE, F. M. H. C. R. Deep learning via semi-supervised embedding. **Neural Networks: Tricks of the Trade**, [S.l.], p.639–655, 2012.

YU D.; SELTZER, M. L. L. J. H. J. S. F. Feature learning in deep neural networks—A study on speech recognition tasks. **CoRR**, [S.l.], v.abs/1301.3605, p.1–9, 2013.

ZOVICO, N. A Tecnologia evoluiu muito! **Momento surdo, Revista Nacional de Reabilitação – REAÇÃO**, [S.l.], v.XV, n.85, p.30, mar-abr 2012.

APÊNDICE A - Coleta inicial para definição e validação da proposta

Pesquisa sobre T ecnologia Assistiva para emissão de ALERTAS.

Prezado (a).
Você é convidado(a) para participar desta pesquisa para levantamento de dados relacionada à fase de concepção(ideia) de uma tecnologia assistiva para surdos. Esta tecnologia terá o propósito de oferecer informações relacionados a emissões de sons que estejam ocorrendo próximo ao deficiente, em especial, alertas em situação de emergência.

O levantamento é parte integrante da minha dissertação de mestrado na Engenharia de Computação (PPGCOMP) do Centro de Ciências Computacionais (C3) na Universidade Federal do Rio Grande (FURG), sob a orientação da Profa. Dra. Regina Barwaldt.

Todos os dados informados serão tratados com total confidencialidade pelo pesquisador, os resultados serão apresentados de forma global para análise.

Sua participação é muito importante para o sucesso dessa pesquisa!

Muito Obrigado,
Douglas Severo Silveira

* Required

1. Marque as alternativas que correspondem sua relação com a surdez: *
   Check all that apply.
   - [ ] Surdo
   - [ ] Tem perda auditiva
   - [ ] Tem familiar(es) surdo(s)
   - [ ] Professor de LIBRAS
   - [ ] Intérprete
   - [ ] Outra relação com a área (indique abaixo)
   - [ ] Other: _____

2. Se possui outra relação com a surdez, informe.
   _____

3. Você considera viável a apresentação de ALERT A VISUAL por meio de reconhecimento de SOM em ambiente externo ou controlado? *
   Mark only one oval.
   - ◯ Sim
   - ◯ Não
   - ◯ Não sei

4. Quais alertas, você imagina, seriam importantes que a tecnologia informasse aos surdos? LISTE POR ORDEM DE IMPORTÂNCIA *
   _____
   _____
   _____
   _____
   _____

Observe a imagem abaixo. Ela representa a forma de funcionamento de nossa proposta.

Sons emitidos são reconhecidos por um smartphone que gera um alerta visual sugerindo o que está acontecendo.

Destes alertas abaixo, quais seriam mais relevantes para os surdos?

5. *
Check all that apply.

- [ ] Veículo dando ré
- [ ] Cachorro latindo
- [ ] Sinal para ônibus parar
- [ ] Buzina de carro
- [ ] Sirene de polícia
- [ ] Sirene de ambulância
- [ ] Sirene de bombeiros
- [ ] Campainha de casa
- [ ] Alguém batendo na porta de casa
- [ ] Sirene te alerta para incêndio
- [ ] Bebê chorando

Agora que já pensamos sobre esta ideia, queremos saber se ela seria importante para você.

6. Você conhece algum recurso com a mesma funcionalidade da apresentada anteriormente? Qual(is)? *

_____

7. Na sua opinião este projeto: *
Mark only one oval.

- ◯ É uma NECESSIDADE e CONTRIBUIRIA para proporcionar ou ampliar habilidades de surdos e, consequentemente, promover Vida Independente e Inclusão.
- ◯ NÃO é uma necessidade, mas CONTRIBUIRIA à vida independente e inclusão de surdos
- ◯ NÃO é uma necessidade e NÃO contribuiria para vida independente e inclusão de surdos

8. Alguma sugestão ou crítica? *

_____

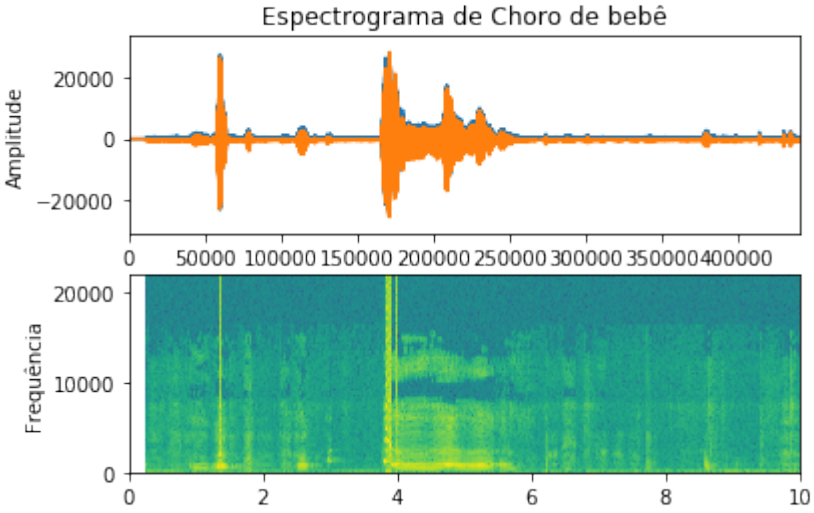
_____

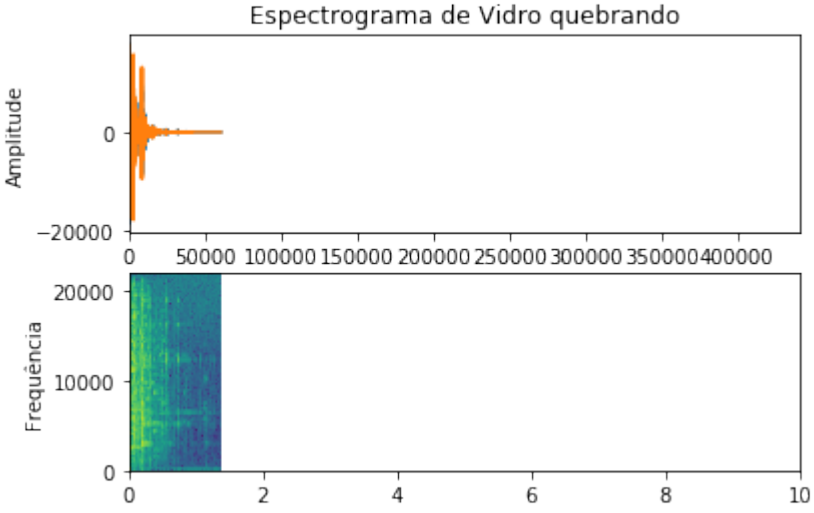
_____

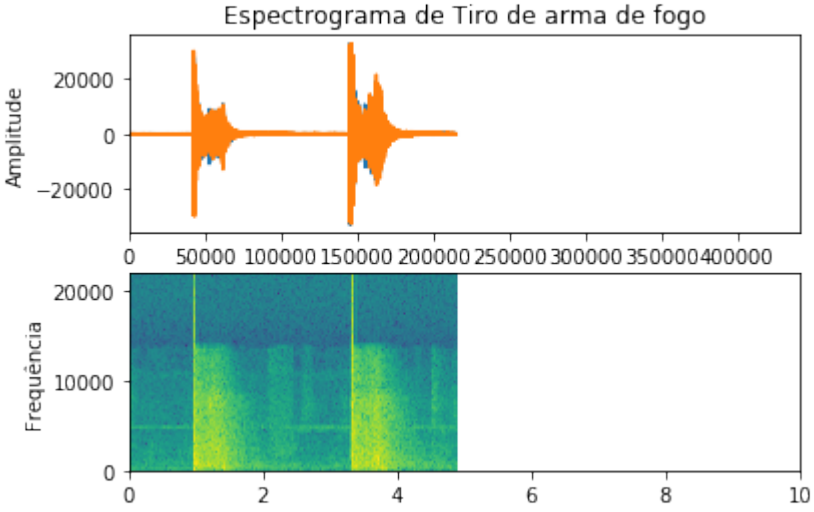
_____

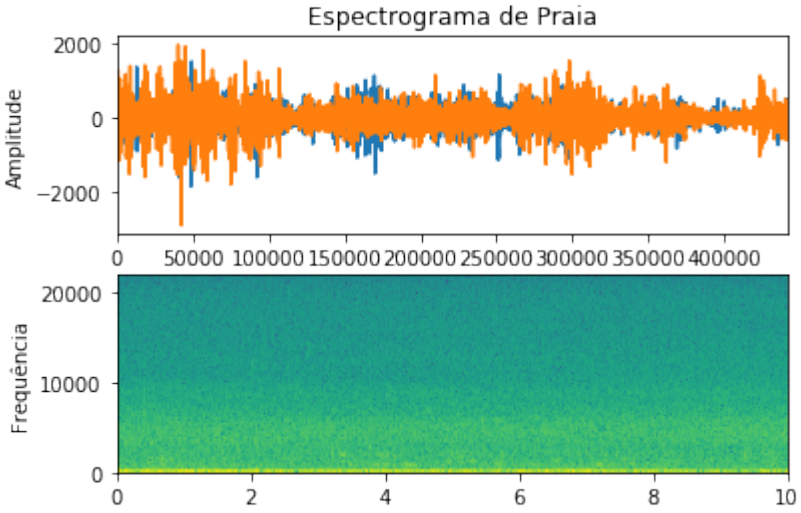
_____

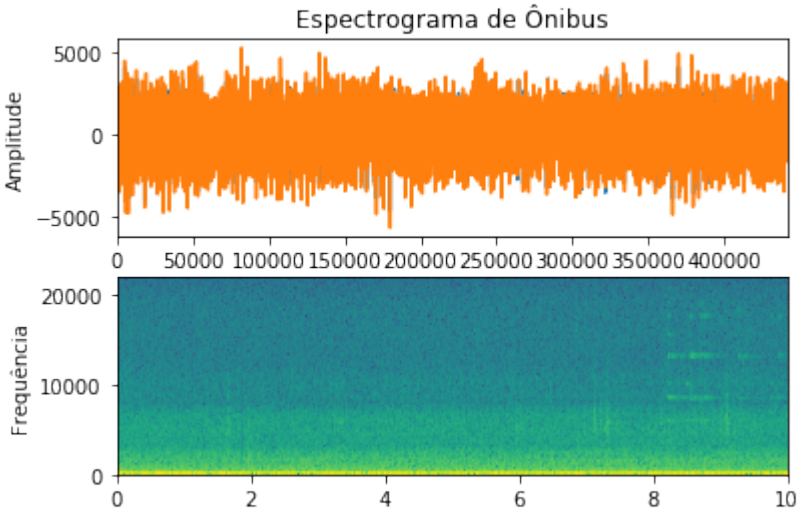
APÊNDICE B - GRÁFICOS EM FORMA DE ONDA E ESPECTROGRAMAS DE SONS DE EVENTO E AMBIENTES

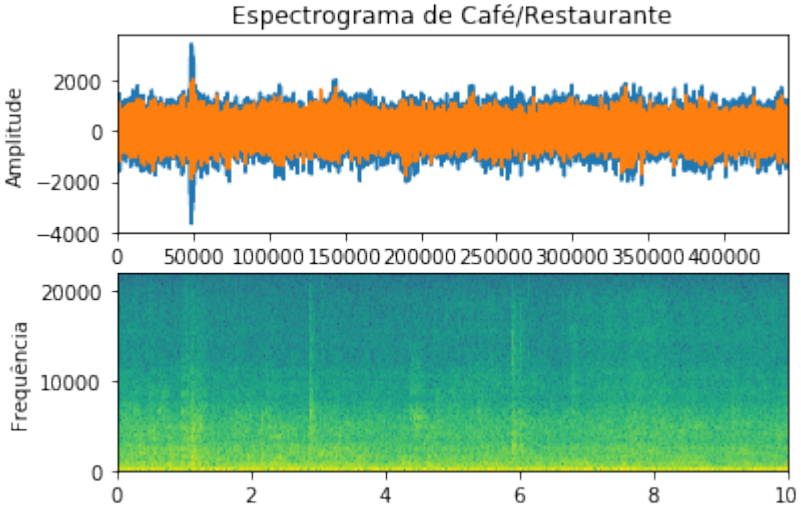

Espectrograma de Choro de bebê
Amplitude
20000
0
−20000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

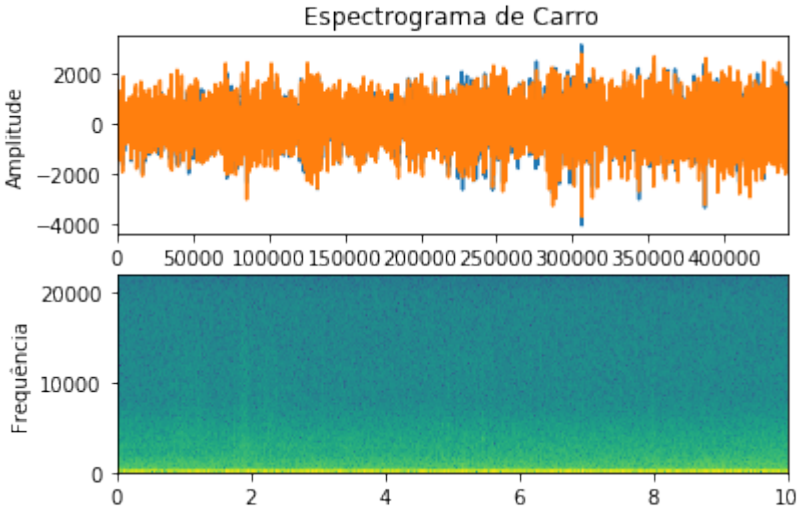

Espectrograma de Vidro quebrando
Amplitude
0
−20000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

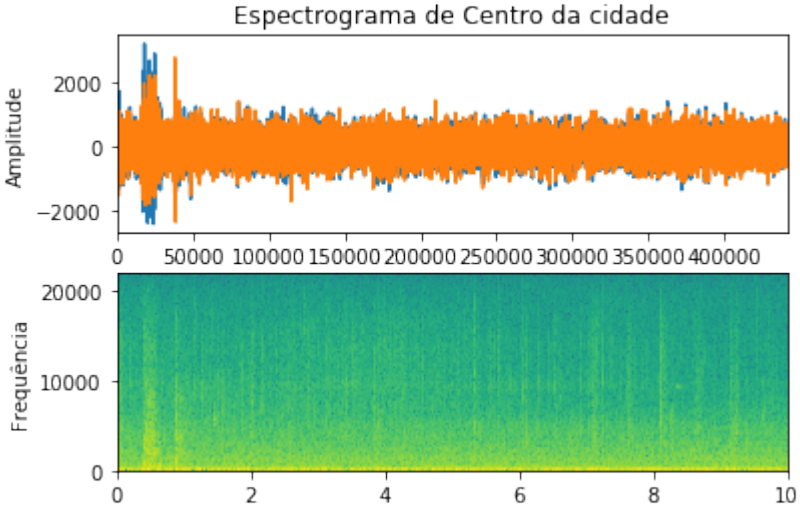

Espectrograma de Tiro de arma de fogo
Amplitude
20000
0
−20000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

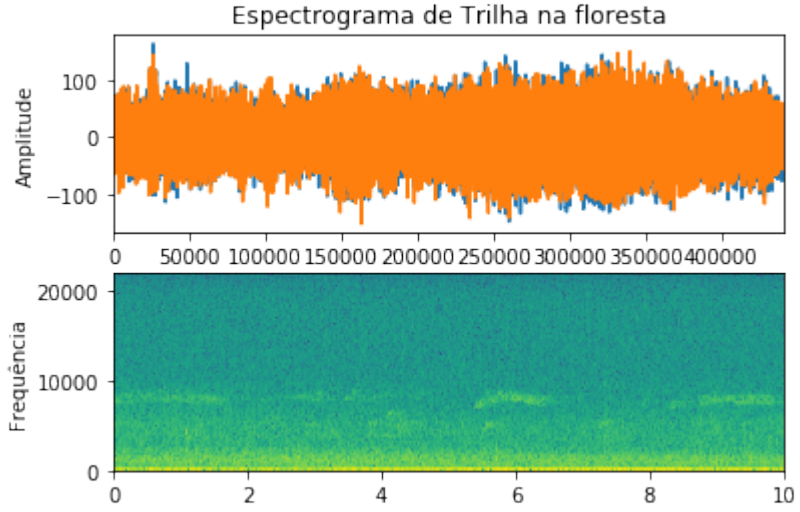

Espectrograma de Praia
Amplitude
2000
0
−2000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

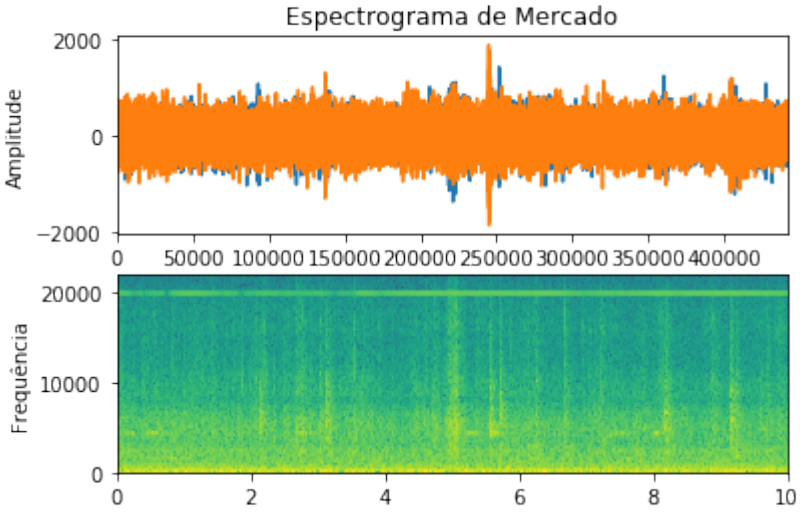

Espectrograma de Ônibus
Amplitude
5000
0
−5000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

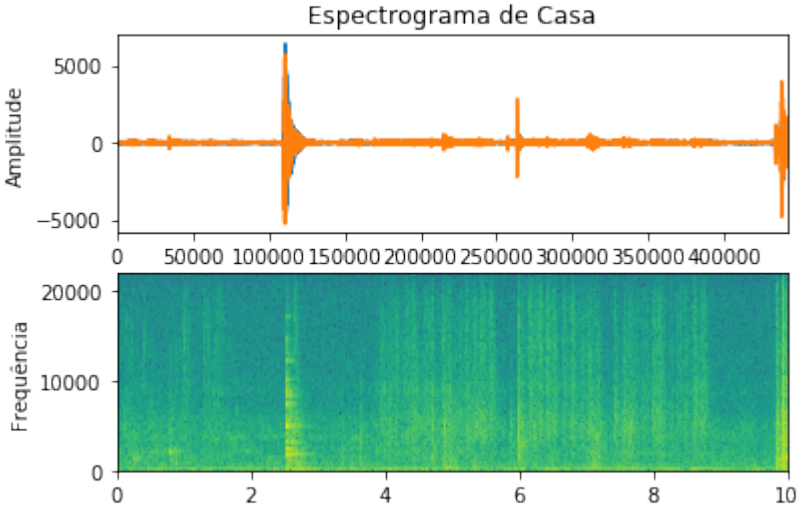

Espectrograma de Café/Restaurante
Amplitude
2000
0
−2000
−4000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

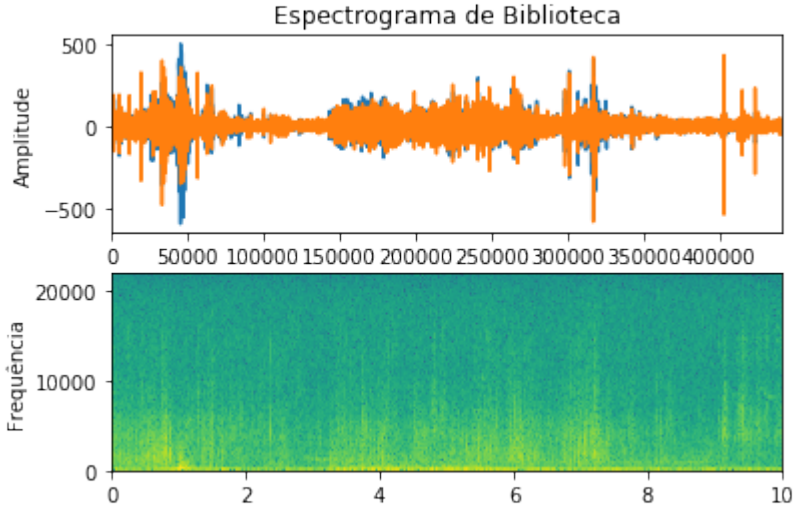

Espectrograma de Carro
Amplitude
2000
0
−2000
−4000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

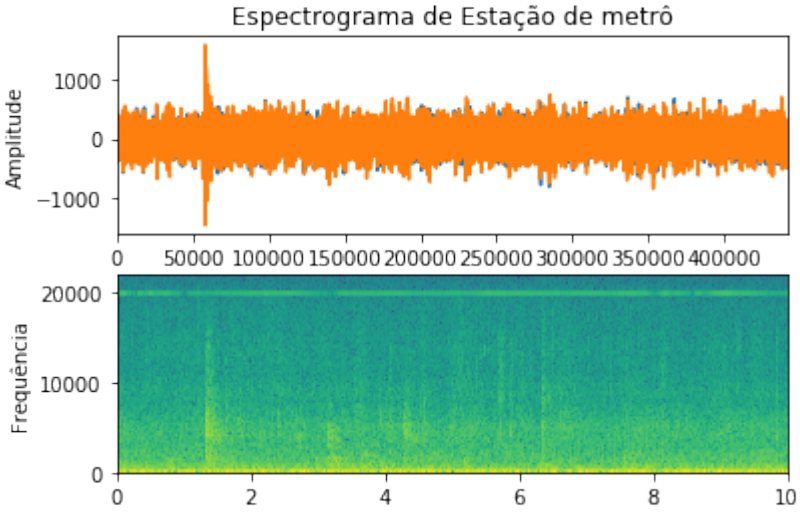

Espectrograma de Centro da cidade
Amplitude
2000
0
−2000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

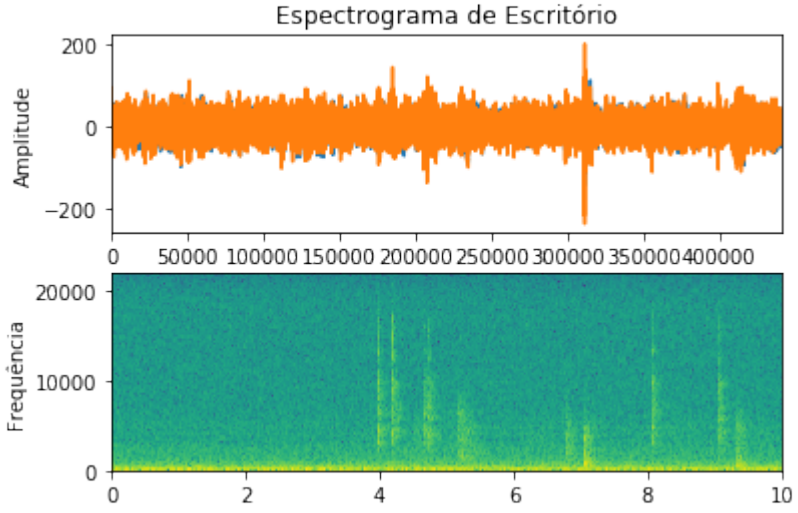

Espectrograma de Trilha na floresta
Amplitude
100
0
−100
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

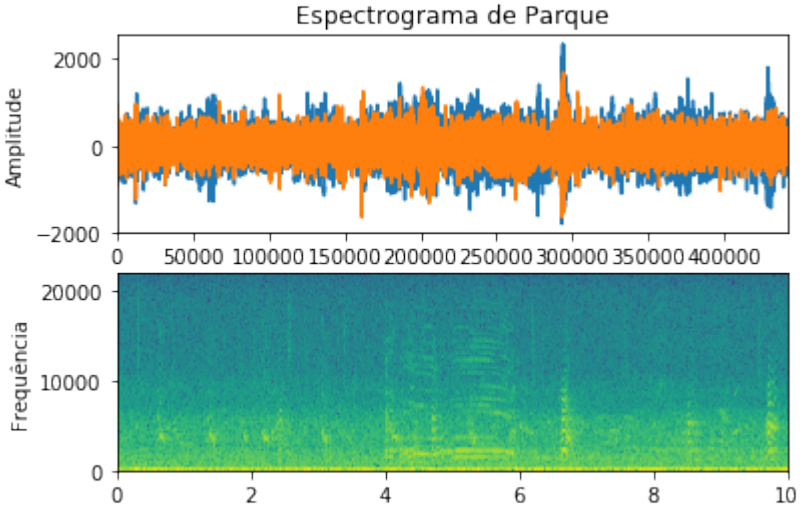

Espectrograma de Mercado
Amplitude
2000
0
−2000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10

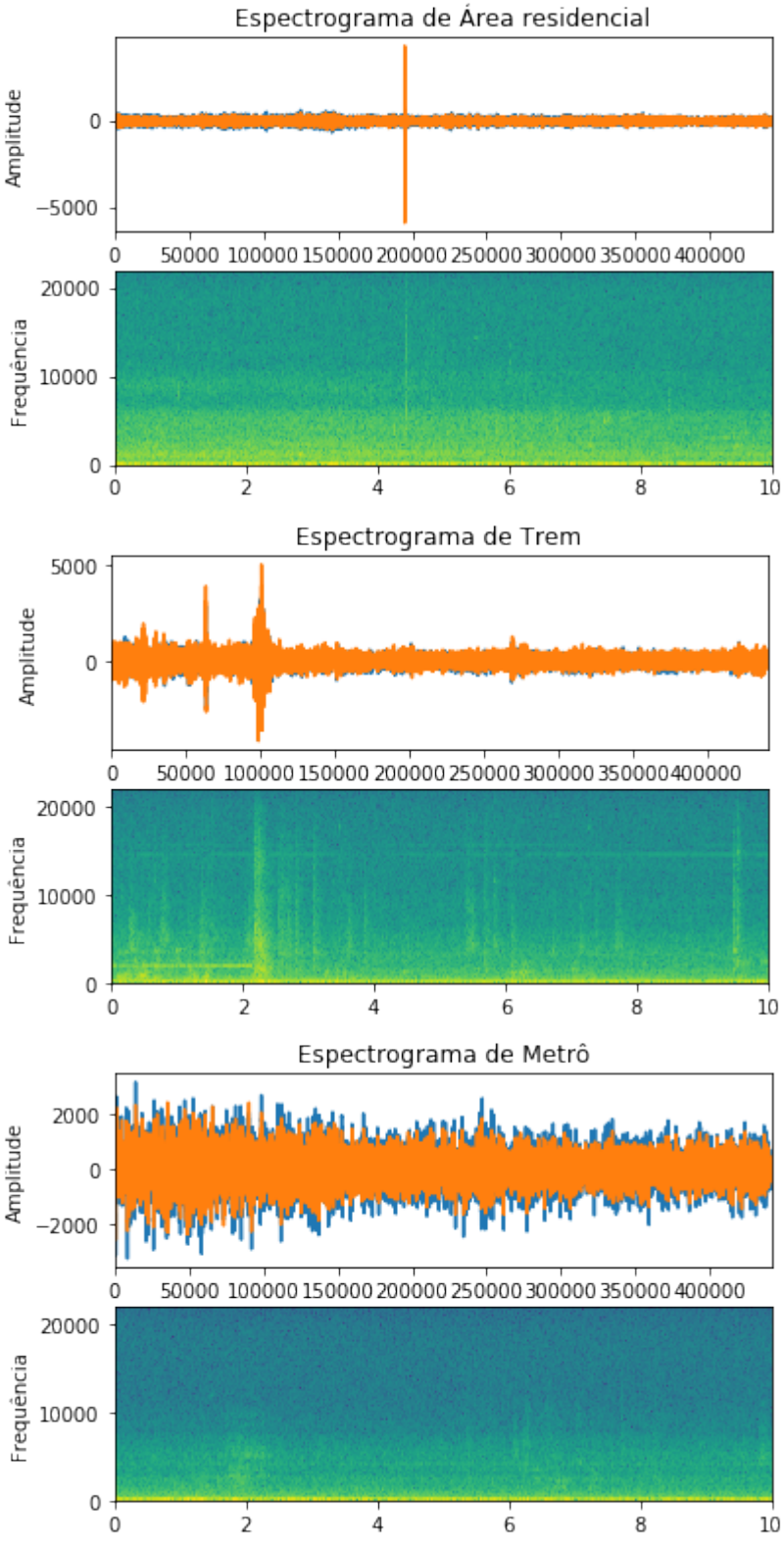

Espectrograma de Casa
Amplitude
5000
0
−5000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10


Espectrograma de Biblioteca
Amplitude
500
0
−500
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10


Espectrograma de Estação de metrô
Amplitude
1000
0
−1000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10


Espectrograma de Escritório
Amplitude
200
0
−200
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10


Espectrograma de Parque
Amplitude
2000
0
−2000
0
50000
100000
150000
200000
250000
300000
350000
400000
Frequência
20000
10000
0
0
2
4
6
8
10


Espectrograma de Área residencial
Amplitude
0
−5000
0 50000 100000 150000 200000 250000 300000 350000 400000
Frequência
20000
10000
0
0 2 4 6 8 10
Espectrograma de Trem
Amplitude
5000
0
0 50000 100000 150000 200000 250000 300000 350000 400000
Frequência
20000
10000
0
0 2 4 6 8 10
Espectrograma de Metrô
Amplitude
2000
0
−2000
0 50000 100000 150000 200000 250000 300000 350000 400000
Frequência
20000
10000
0
0 2 4 6 8 10

APÊNDICE C - CONFIGURAÇÃO REDE NEURAL CONV1

```
inp = Input(shape=(input_length,1))
x = Convolution1D(16, 9, activation=relu, padding="valid")(inp)
x = Convolution1D(16, 9, activation=relu, padding="valid")(x)
x = MaxPool1D(16)(x)
x = Dropout(rate=0.1)(x)

x = Convolution1D(32, 3, activation=relu, padding="valid")(x)
x = Convolution1D(32, 3, activation=relu, padding="valid")(x)
x = MaxPool1D(4)(x)
x = Dropout(rate=0.1)(x)

x = Convolution1D(32, 3, activation=relu, padding="valid")(x)
x = Convolution1D(32, 3, activation=relu, padding="valid")(x)
x = MaxPool1D(4)(x)
x = Dropout(rate=0.1)(x)

x = Convolution1D(256, 3, activation=relu, padding="valid")(x)
x = Convolution1D(256, 3, activation=relu, padding="valid")(x)
x = GlobalMaxPool1D()(x)
x = Dropout(rate=0.2)(x)

x = Dense(64, activation=relu)(x)
x = Dense(1028, activation=relu)(x)
out = Dense(nclass, activation=softmax)(x) #nclass: quantidade de
classes
```

APÊNDICE D - CONFIGURAÇÃO REDE NEURAL CONV2

```
inp = Input(shape=(config.dim[0],config.dim[1],1))
x = Convolution2D(32, (4,10), padding="same")(inp)
x = BatchNormalization()(x)
x = Activation("relu")(x)
x = MaxPool2D()(x)

x = Convolution2D(32, (4,10), padding="same")(x)
x = BatchNormalization()(x)
x = Activation("relu")(x)
x = MaxPool2D()(x)

x = Convolution2D(32, (4,10), padding="same")(x)
x = BatchNormalization()(x)
x = Activation("relu")(x)
x = MaxPool2D()(x)

x = Convolution2D(32, (4,10), padding="same")(x)
x = BatchNormalization()(x)
x = Activation("relu")(x)
x = MaxPool2D()(x)

x = Flatten()(x)
x = Dense(64)(x)
x = BatchNormalization()(x)
x = Activation("relu")(x)
out = Dense(nclass, activation=softmax)(x) #nclass: quantidade de
classes
```