

**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO DE ENGENHARIA DE CONTROLE E AUTOMAÇÃO**

MAURICIO YOITI ISHIZAKI

RECONHECIMENTO AUTOMÁTICO DE PALAVRAS

TRABALHO DE CONCLUSÃO DE CURSO

**CORNÉLIO PROCÓPIO
2018**

MAURICIO YOITI ISHIZAKI

RECONHECIMENTO AUTOMÁTICO DE PALAVRAS

Trabalho de Conclusão de Curso de graduação, apresentado à disciplina Trabalho de conclusão de curso, do curso de Engenharia de controle e automação da Universidade Tecnológica Federal do Paraná – UTFPR, como requisito parcial para a obtenção do título de Engenheiro.

Orientador: Prof. Dra. Maria Eugenia Dajer
Coorientador: Prof. Dr. Danilo Hernane Spatti

CORNÉLIO PROCÓPIO
2018



Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio
Departamento Acadêmico de Elétrica
Curso de Engenharia de Controle e Automação



FOLHA DE APROVAÇÃO

Mauricio Yoiti Ishizaki

Reconhecimento automático de palavras

Trabalho de conclusão de curso apresentado às 15:50hs do dia 19/06/2018 como requisito parcial para a obtenção do título de Engenheiro de Controle e Automação no programa de Graduação em Engenharia de Controle e Automação da Universidade Tecnológica Federal do Paraná. O candidato foi arguido pela Banca Avaliadora composta pelos professores abaixo assinados. Após deliberação, a Banca Avaliadora considerou o trabalho aprovado.

Prof(a). Dr(a). **María Eugenia Dajer** - Presidente (Orientador)

Prof(a). Dr(a). **Danilo Hernane Spatti** - (Coorientador)

Prof(a). Dr(a). **Cristiano Marcos Agulhari** - (Membro)

Prof(a). Dr(a). **Alessandro Goedel** - (Membro)

A folha de aprovação assinada encontra-se na coordenação do curso.

AGRADECIMENTOS

Aos meus pais e meu irmão, por todo o carinho e apoio incondicional em meus projetos e sonhos. Sempre acreditando em mim e me incentivando a aprender sempre.

À minha namorada, pelo carinho, paciência e apoio. Mesmo estando compromissada com sua graduação e estágio, me ajudou em vários momentos da elaboração deste trabalho.

Aos meus orientadores, Prof. Dr. Danilo Hernane Spatti e Prof. Dra. Maria Eugenia Dajer, pela contribuição e orientação durante toda a elaboração deste trabalho e pelo companheirismo, bom humor e confiança.

A todos os amigos que estiveram comigo durante esta jornada e que fizeram parte da minha vida.

À Universidade Tecnológica Federal do Paraná, todos os professores e funcionários que contribuíram para a minha formação.

RESUMO

ISHIZAKI, Mauricio Yoiti. Reconhecimento automático de palavras. 2018. 43f. Trabalho de Conclusão de curso (Graduação) – Engenharia de Controle e Automação. Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2018.

Reconhecimento automático de palavras é a tradução da fala humana para texto, o qual se mostrou útil na comunicação homem - computador. Por este motivo, várias pesquisas foram feitas nesta área e conseqüentemente aplicações, como assistentes virtuais, surgiram para facilitar a vida das pessoas. No entanto, pessoas que sofrem com alguma disfonia (rouquidão) não conseguem desfrutar totalmente destas aplicações, devido às distorções da voz. Este trabalho propõe a utilização de Redes Neurais Convolucionais (CNNs) para fazer o reconhecimento de palavras faladas com esse tipo de distorção. Utilizou-se uma base de dados de 20 palavras com 28 amostras, sendo todas as vozes de diferentes pessoas disfônicas. Foram criadas diversas topologias para a CNN, variando alguns hiperparâmetros da rede. Foi feito o treinamento e teste de cada uma dela. Para o conjunto de teste, a topologia com maior acurácia obteve um resultado de 82,50%.

Palavras-chave: Reconhecimento automático de palavras, Rede Neural Profunda, Rede neural convolucional.

ABSTRACT

ISHIZAKI, Mauricio Yoiti. Automatic extraction of words. 2018 43f. Course Completion Work (Graduation) - Control and Automation Engineering. Federal Technological University of Paraná. Cornélio Procópio, 2018.

Automatic word recognition is the translation of human speech into text, which has proved useful in man - computer communication. For this reason, several researches were developed in this area and consequently applications, such as virtual assistants, have arisen to make life easier for people. However, people who suffer from any dysphonia (hoarseness) can't fully enjoy these applications, due to the distortions in their voice. This paper proposes the use of Convolutional Neural Networks (CNNs) to make the recognition of spoken words with this type of distortion. A database of 20 words with 28 samples was used, all voices were from different dysphonic people. Several topology were created for CNN, varying some hyperparameters of the network. All topologies were training and testing. For the test set, the topology with the highest accuracy obtained a result of 82,50%.

Keywords: Automatic word recognition, Deep Neural Network, Convolutional Neural Network

LISTA DE ILUSTRAÇÕES

Figura 1 – Representação de uma DNN.....	17
Figura 2 – Representação do campo receptivo local em CNN.....	19
Figura 3 – Exemplo de Convolução unidimensional de um mapa de características	20
Figura 4 – Exemplo de compartilhamento de pesos (w) e limiares (b).....	21
Figura 5 – Exemplo de aplicação do <i>ReLU</i> e <i>pooling</i>	22
Figura 6 – Diagrama de representação de uma rede neural convolucional.....	22
Figura 7 – Exemplo de aplicação de <i>dropout</i> em uma rede neural convencional	24
Figura 8 – Diagrama de blocos de um sistema de reconhecimento de voz	27
Figura 9 – Diagrama de blocos da fase de pré-processamento.....	27
Figura 10 – Espectrograma da palavra papai.....	28
Figura 11 – Arquitetura completa para o modelo CNN proposto.....	30
Figura 12 – Porcentagem de acerto para cada palavra.....	34

LISTA DE QUADROS

Quadro 1 – Número de citações realizadas utilizando aprendizagem profunda para reconhecimento de fala.....	18
Quadro 2 – Número de amostras de cada palavra.....	26
Quadro 3 – Acurácia das amostras de treinamento, validação e teste para cada configuração da CNN.....	33
Quadro 4 – Matriz de confusão do resultado com maior acurácia	36

LISTA DE SIGLAS

ML	Aprendizado de máquina (Machine Learning)
DNN	Rede Neural Profunda (Deep Neural Network)
CNN	Rede Neural Convolucional (Convolutional Neural Network)
Matlab	Matrix Laboratory

SUMÁRIO

1	INTRODUÇÃO.....	10
1.1	Problemas e premissas.....	10
1.2	Justificativas.....	11
1.3	Objetivos.....	11
1.3.1	Objetivo geral.....	11
1.3.2	Objetivos específicos.....	12
2	FUNDAMENTAÇÃO TEÓRICA.....	13
2.1	Reconhecimento de fala.....	13
2.1.1	Classificação do reconhecimento de fala	13
2.1.2	Aplicações do reconhecimento de fala.....	14
2.2	Aprendizagem de máquina.....	15
2.2.1	Aprendizagem Profunda.....	16
2.2.1.1	Rede Neural Convolutacional.....	18
3	MATERIAIS E MÉTODOS.....	25
3.1	Materiais.....	25
3.2	Métodos.....	26
3.2.1	Pré-processamento.....	27
3.2.2	Extração de características do sinal da fala.....	29
3.2.3	Reconhecimento.....	31
4	RESULTADOS E DISCUSSÕES.....	33
5	CONSIDERAÇÕES FINAIS.....	37
5.1	Trabalhos Futuros.....	37
6	REFERÊNCIAS.....	39

1. INTRODUÇÃO

1.1 Problemas e premissas

Reconhecimento automático de palavras, de acordo com Stuckless (1994), pode ser definido como a transcrição de linguagem falada para texto legível em tempo real. Em outras palavras, é a tecnologia que permite que um computador gere um texto escrito a partir do reconhecimento de palavras ditas por uma pessoa.

Desenvolver sistemas que possam realizar este tipo de tarefa conduziram pesquisas por mais de 50 anos e, graças a estas pesquisas, o reconhecimento automático de palavras apresentou progressos significativos (COLE; et al, 1996), (HINTON; et al, 2012), (ABDEL-HAMID; et al, 2013). Conseqüentemente diversas aplicações foram desenvolvidas, tais como interfaces de sistemas eletrônicos, controle de ambientes, telemarketing e até mesmo auxílio a deficientes. Porém, devido às características do sinal de voz, ainda existe uma lacuna entre reconhecimento de fala humana e o feito por computadores (GARG; SHARMA, 2016).

A complexidade dos sistemas de reconhecimento de voz é elevada, pois a voz humana é composta por interações de diversos órgãos e, como o organismo de cada indivíduo possui características únicas, a fala de duas pessoas nunca é exatamente a mesma. Isso se deve em parte a quatro fontes de variabilidade associadas ao sinal de voz (COLE; et al, 1996):

A primeira delas é a variabilidade fonética, na qual a pronúncia dos fonemas são altamente dependentes do contexto em que aparecem, como por exemplo o fonema /x/ em xícara, texto e táxi, no português brasileiro. Outra característica é que nos limites das palavras, as variações contextuais podem ser bastante dramáticas. Por exemplo a frase 'ela alimenta o ...' pode ser pronunciada como 'ela alimento...'

A segunda fonte de variabilidade advém de mudanças no ambiente, assim como da posição e características do transdutor. A terceira configura-se como variabilidades intra-locutor, resultante de mudanças do estado físico/emocional dos locutores, velocidade de pronúncia ou qualidade de voz. Por fim, a quarta são as variabilidades entre-locutores que é originada das diferenças na condição sócio-cultural, dialeto, tamanho e forma do trato vocal de cada pessoa.

1.2 Justificativa

Com o avanço da qualidade dos sistemas de reconhecimento de fala e, conseqüentemente, o desenvolvimento de diversas aplicações, tais sistemas começaram a fazer parte do cotidiano das pessoas, em aplicativos de smartphones e em assistentes virtuais em computadores. Isso ocorre, pois, além da facilidade e conforto para a vida do ser humano, também pode servir de ajuda a deficientes físicos, como por exemplo escrever uma frase por meio do uso de sua fala ou realizar o acionamento de uma máquina para pessoas que não possam se locomover para ativar manualmente.

Pessoas que sofrem com patologias na voz, no entanto, tem dificuldade ao utilizarem sistemas de reconhecimento de voz, pois as perturbações presentes na voz destas afetam negativamente no reconhecimento da fala, como pode ser observado em (VACHER et al., 2015).

Uma alternativa para resolver este problema é pela utilização de redes neurais convolucionais. Isso se deve ao fato que esse tipo de rede foi projetada para lidar com variabilidade no deslocamento, escala e distorção (LECUN et al., 1998). Portanto, a motivação e proposta deste trabalho é avaliar a utilização de topologias de redes neurais convolucionais para fazer o reconhecimento de palavras faladas por pessoas que sofrem com patologias que geram disfonia, de forma eficiente e independente de locutor.

1.3 OBJETIVOS

1.3.1 Objetivo geral

Realizar o reconhecimento automático de palavras pronunciadas por pessoas com alguma disfonia, utilizando redes neurais convolucionais.

1.3.2 Objetivos específicos

- Preparar o banco de dados para desenvolvimento do sistema de reconhecimento de fala;
- Diminuir o sobre-ajuste (*overfitting*) do modelo proposto;
- Verificar, implementar e validar o CNN utilizado;
- Analisar e avaliar resultados obtidos.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Reconhecimento de fala

De acordo com Dahl (2012), reconhecimento da fala é o processo de converter um sinal acústico, capturado por um transdutor como um microfone ou um telefone, para um conjunto de palavras.

O sinal de fala é obtido pela conversão de ondas sonoras em sinais elétricos por meio de um transdutor, sendo o microfone o mais comumente utilizado. É feita então a filtragem desse sinal por meio de um filtro *anti-aliasing*, sendo este necessário para atenuar as componentes de frequência superiores à metade da frequência de amostragem a ser utilizada, atendendo o critério de Nyquist (PROAKIS; MANOLAKIS, 1996).

Utiliza-se, então, um conversor analógico-digital (A/D) para converter o sinal de fala analógico em digital (SILVA, 2009). A partir desse sinal, realiza-se o reconhecimento de fala para obter o conjunto de palavras.

As palavras reconhecidas podem ser o resultado final do sistema como no caso de aplicações de comandos de controle ou servir de entrada de dados e documentos de preparação para outros sistemas (DAHL, 2012).

2.1.1 Classificação do reconhecimento de fala

Um sistema de reconhecimento de fala pode ser classificado de acordo com o modo de pronúncia, podendo ele ser dividido em três tipos: Reconhecimento de palavras isoladas, palavras conectadas e fala contínua.

No reconhecimento de palavras isoladas, são reconhecidas as palavras pronunciadas isoladamente ou que possuam uma pausa mínima entre elas. No reconhecimento de palavras conectadas, sentenças completas podem ser reconhecidas. Já para o reconhecimento de fala contínua, o sistema deve ser capaz de reconhecer a comunicação natural (SILVA, 2009).

Reconhecimento de fala também pode ser classificado como dependente ou não do locutor. Os sistemas dependentes de locutor realizam o reconhecimento da fala de um único locutor ao qual foi treinado. Os independentes, por sua vez, são mais

sofisticados, pois são aqueles capazes de identificar uma entrada falada por diferentes locutores, dado um vocabulário específico, sem a necessidade de um treinamento prévio (VALIATI, 2000).

Para a classificação dos reconhecedores de fala, o tamanho do vocabulário é um outro aspecto importante que deve ser considerado. O tamanho é separado em quatro categorias (INCE,2013):

- Vocabulário Pequeno: 1 a 20 palavras.
- Vocabulário Médio: 20 a 100 palavras.
- Vocabulário Grande: 100 a 1000 palavras.
- Vocabulário Muito Grande: mais de 1000 palavras.

2.1.2 Aplicações do reconhecimento de fala

Os sistemas de reconhecimento de fala são extremamente práticos e úteis para a vida do ser humano atualmente, por isso podem ser utilizados em diversas atividades (VALIATI, 2000):

- Interface em sistemas eletrônicos: O reconhecimento de fala, neste tipo de aplicação, funciona como uma interface entre usuário e aparelhos – como computadores e celulares – no qual a fala faz o papel de manipular o sistema e seus aplicativos. Como exemplos de aplicação podem-se citar abrir e fechar aplicativos, fazer pesquisas na internet, tocar músicas, tirar fotos com a câmera do celular e diversas outras ações.

- Auxílio a deficientes: O reconhecimento de fala também pode ser utilizado para auxiliar pessoas com alguma deficiência física e com problemas motores, como pessoas paraplégicas ou com Síndrome do Esforço Repetitivo, fazendo o acionamento de equipamentos e dispositivos que possam ajudá-los em seu cotidiano. Essa aplicação também é importante para indivíduos com deficiência visual, pois pode captar e transcrever textos falados por estes, e deficientes auditivos, criando sistemas de conversação telefônica na qual o que é falado ao telefone é traduzido para o sistema Braille.

- Controle de ambientes: Outra aplicação é o controle de ambientes domiciliares como sala, cozinha, quarto e banheiro. A fala pode ser utilizada para ligar luzes, ligar ventiladores ou equipamentos de ar-condicionado, aparelhos televisores e

quaisquer outros equipamentos que tenham sido programados, tornando o ambiente muito mais aconchegante para as necessidades de cada usuário. Tal aplicação pode ser feita em ambientes hospitalares fazendo o ajuste das camas, clima do ambiente e até na movimentação de cadeiras de rodas.

- **Telecomunicações:** Este é um campo que pode ser usado como serviço de discagem pela voz e direcionamento de chamadas apenas dizendo um número de identificação ou somente o apelido da pessoa. O destaque destes serviços é a utilização de páginas amarelas, ou seja, o usuário fala o nome de uma empresa, produto ou nome fantasia e o sistema retorna o número ou até mesmo faz a ligação direta com a empresa.

2.2 Aprendizado de máquina

Para Mitchell (1997) a definição de aprendizado de máquina, do inglês *machine learning* (ML), é dada como: "Se diz que um programa de computador aprende pela experiência E com respeito a algum tipo de tarefa T e performance P se sua performance P nas tarefas em T , como medidas por P , melhoram com a experiência E ."

De acordo com Bishop (2006) e Mehryar; Rostamizadeh; Talwalkar (2012), o objetivo do ML é generalizar as saídas do sistema a partir de suas experiências, ou seja, após ser apresentada a um conjunto de dados de aprendizado, a máquina deve ser capaz de executar com precisão tarefas ainda não vistas. Nota-se, portanto, que ao contrário de sistemas inteligentes, que são fundamentados na observação biológica, o ML está focado no aprendizado de computadores.

Este método pode ser classificado em aprendizagem supervisionada ou não supervisionada, os quais são diferenciados por conter ou não rótulos (valores discretos esperados) do sistema (HAYKIN, 1997).

Para a aprendizagem supervisionada, também conhecida como aprendizado a partir de exemplos, ocorre a geração de um classificador que, a partir de um conjunto de dados reais de treinamento e da saída do sistema, consegue prever a classe de novos exemplos após a realização do mapeamento das saídas de acordo com as entradas. (MITRA, 2006), (MARSLAND, 2015).

No aprendizado não-supervisionado não se conhece as respostas do sistema, e o ML aprende por meio de um processo conhecido como *clustering*, no qual o

sistema identifica e detecta singularidades entre as amostras do processo, com o objetivo de agrupamento destes (STEPP; RYSZARD, 1986); (DA SILVA; SPATTI; FLAUZINO, 2010).

Dependendo de como o ML processa um exemplo, suas tarefas podem ser descritas de formas diferentes. Dentre as mais comuns, podem-se citar a classificação e a regressão. Na classificação o programa de computador é requisitado a especificar a qual categoria a entrada pertence, ou seja, o sistema deve generalizar, para predizer um valor discreto em novos exemplos. Na regressão as saídas são contínuas ao invés de discretas, diferente da classificação (RUSSELL, 1995); (GOODFELLOW; BENGIO; COURVILLE, 2017).

2.2.1 Aprendizagem profunda

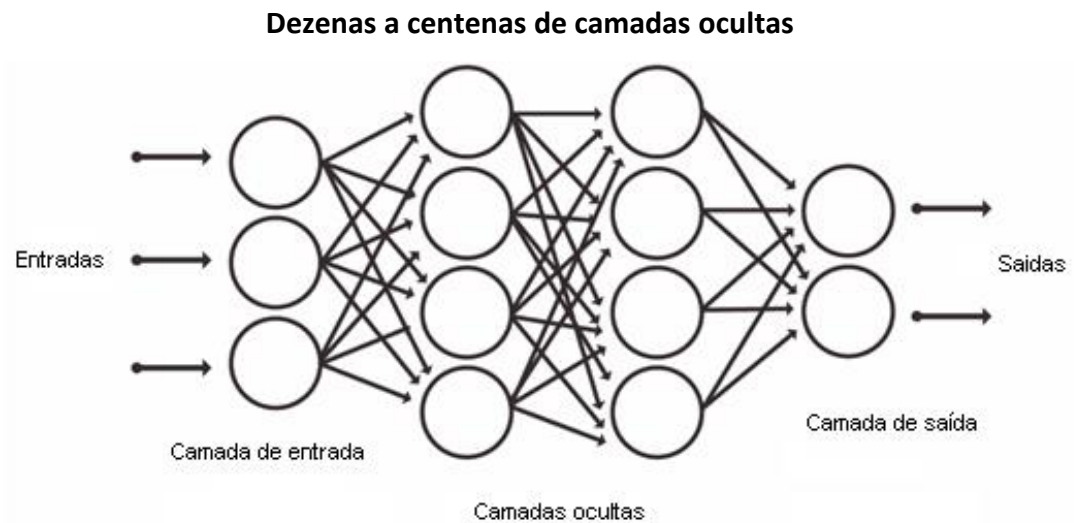
De acordo com Deng; Dong (2014), aprendizagem profunda (*Deep Learning*) é uma classe de técnicas de ML que tem como finalidade realizar o aprendizado de funções não supervisionadas, além de analisar e classificar padrões a partir de várias camadas de processamento não linear de arquiteturas hierárquicas supervisionadas.

Ainda de acordo com Deng; Dong (2014), a ideia fundamental da aprendizagem profunda é computar as características hierárquicas ou a representação dos dados observacionais, na qual características de nível mais alto são compostos por características de níveis inferiores. A extração direta das características dos dados é uma das vantagens mais significativas desta classe.

Frequentemente modelos de aprendizagem profunda são referidos como redes neurais profundas (*deep neural networks* – DNN). Isso se deve ao fato de que a maioria dos métodos de aprendizagem profunda utilizam arquiteturas de redes neurais. O grande número de camadas ocultas presente em uma rede neural motiva a terminologia de “profundo”. A efeito de comparação, enquanto redes neurais tradicionais possuem de 2 a 3 camadas ocultas, algumas das mais recentes redes profundas têm até 150 camadas (PATEL, 2017).

Uma DNN é ilustrada pela Figura 1.

Figura 1: Representação de uma DNN



Fonte: adaptado de (PATEL, 2017)

A desvantagem da utilização de aprendizagem profunda ao invés de outras técnicas de ML, é que o aprendizado profundo é mais complexo, portanto, necessita de maior poder de processamento e quantidade de dados, além de precisar de um período maior de tempo para treinamento. Em compensação, este método pode ser altamente preciso e não gera a necessidade de entender quais as melhores características que representam o objeto (PINGEL, 2017).

Atualmente, devido ao aumento do poder de processamento dos computadores e maior quantidade de dados acessíveis, o método de aprendizagem profunda, para utilização em reconhecimento de fala, tornou-se popular e elevou o número de pesquisas realizadas. Isto pode ser visto pelo grande número de citações de trabalhos consolidados desta área, mostrado no Quadro 1, o que também é motivação para a execução deste trabalho.

Quadro 1: Número de citações realizadas utilizando aprendizagem profunda para reconhecimento de fala

Titulo	Ano	Citações
Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. (SEIDE; Li; Yu, 2011)	2011	575
Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups (HINTON; et al, 2012)	2012	2643
Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition (MAMEDE; et al., 1978)	2012	1428
Acoustic modeling using deep belief networks (MOHAMED; DAHL; HINTON, 2012)	2012	936
Deep learning (LECUN; BENGIO; HINTON, 2015)	2015	2583

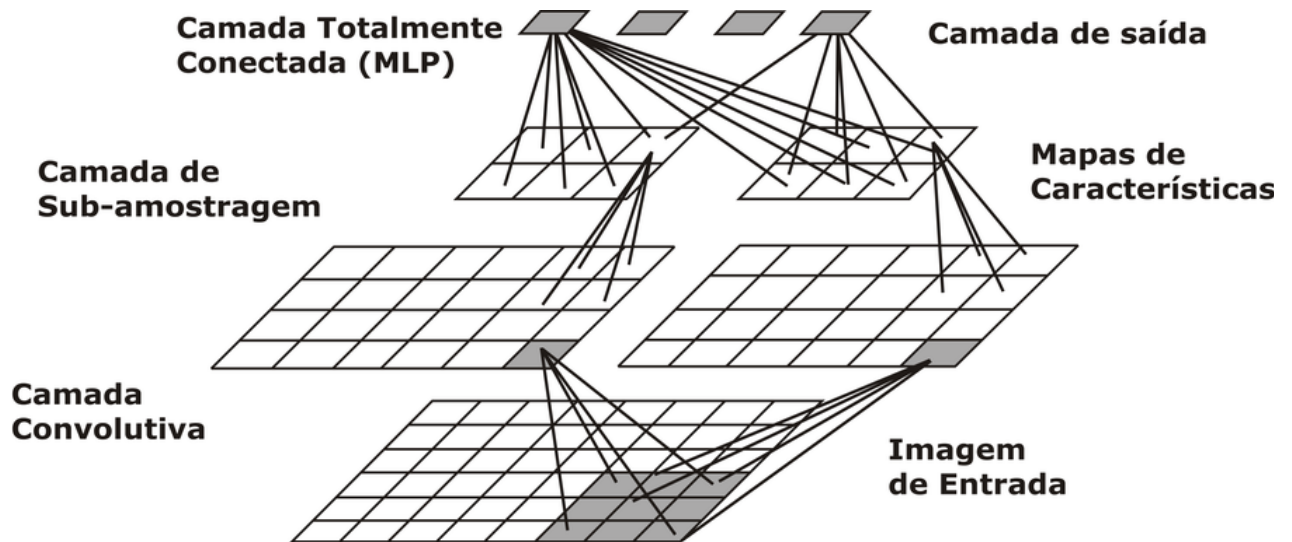
Fonte: autoria própria

2.2.1.1 Rede Neural Convolutacional

Rede neural convolutacional, do inglês *Convolutional Neural Network* (CNN) é um tipo de arquitetura de aprendizagem profunda utilizada para processamento de dados cujas topologias são dadas como matrizes (MARSLAND,2015). A CNN tem como propriedade ser invariante a deslocamento, escala e distorção (LECUN et al., 1998). Para isso, as CNNs combinam três conceitos: campos receptivos locais, pesos e limiares compartilhados, e ativação e *pooling* (PATEL; PINGEL, 2017).

Em CNN, existem regiões conhecidas como campos receptivos locais, que são pequenas regiões da camada oculta conectadas em neurônios da camada de entrada. Essas regiões têm o objetivo de fazer o mapeamento de uma determinada região da matriz de entrada em uma região específica da matriz de saída chamada de mapa de características (*feature maps*) (PATEL; PINGEL, 2017). Isso pode ser visto na Figura 2.

Figura 2: Representação do campo receptivo local e mapa de características de uma CNN.



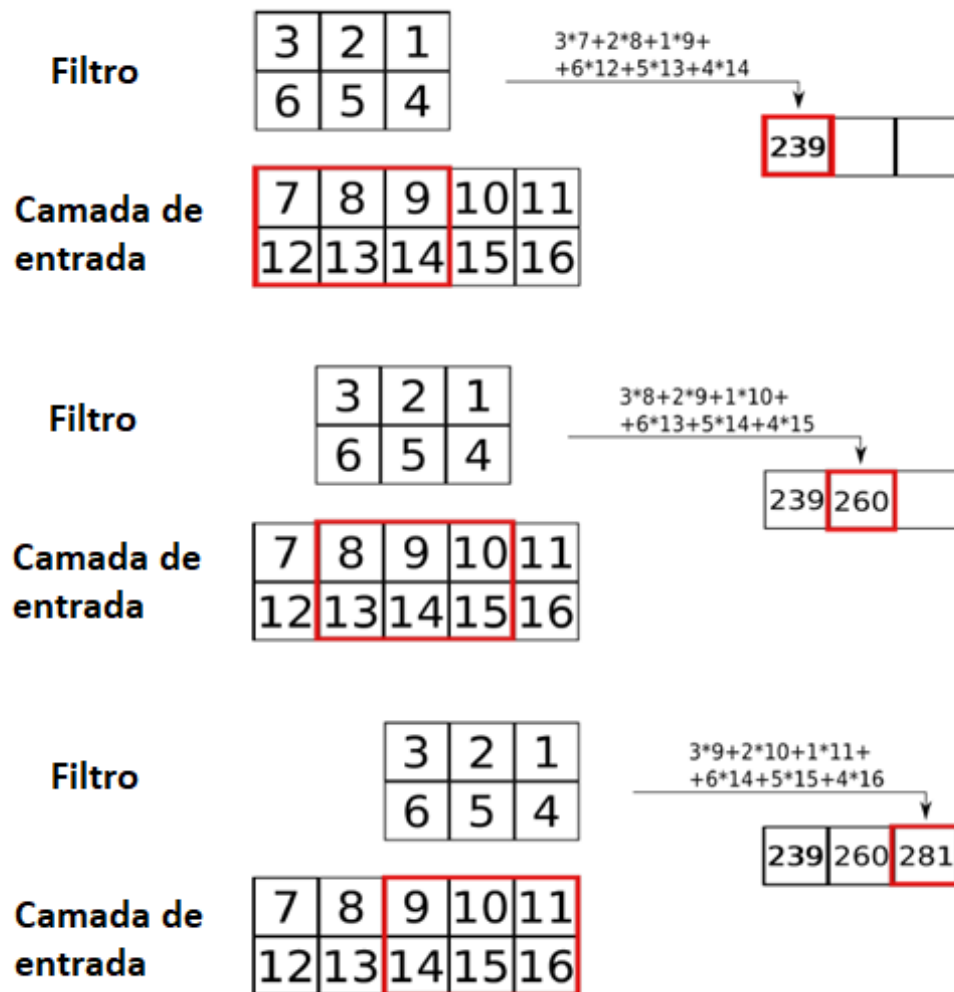
Fonte: Li (1999)

Esses mapas de características são encontrados realizando uma operação equivalente à convolução (LECUN, 1995). Ou seja, realiza-se o cálculo do produto matricial entre um filtro, conhecido como *kernel*, e as entradas (ou camadas anteriores), e por fim soma-se estes resultados.

Para encontrar todos os valores de um mapa de características, desliza-se o mesmo filtro (de acordo com o tamanho do passo escolhido) por toda a camada de entrada e armazenam-se os resultados das convoluções em uma determinada região do mapa de características.

Para a realização destes cálculos, utiliza-se os pesos das conexões entre as camadas de entrada e as camadas ocultas como sendo os filtros das convoluções. Esta implementação pode ser observada na Figura 3.

Figura 3: Exemplo de Convolução unidimensional de um mapa de características

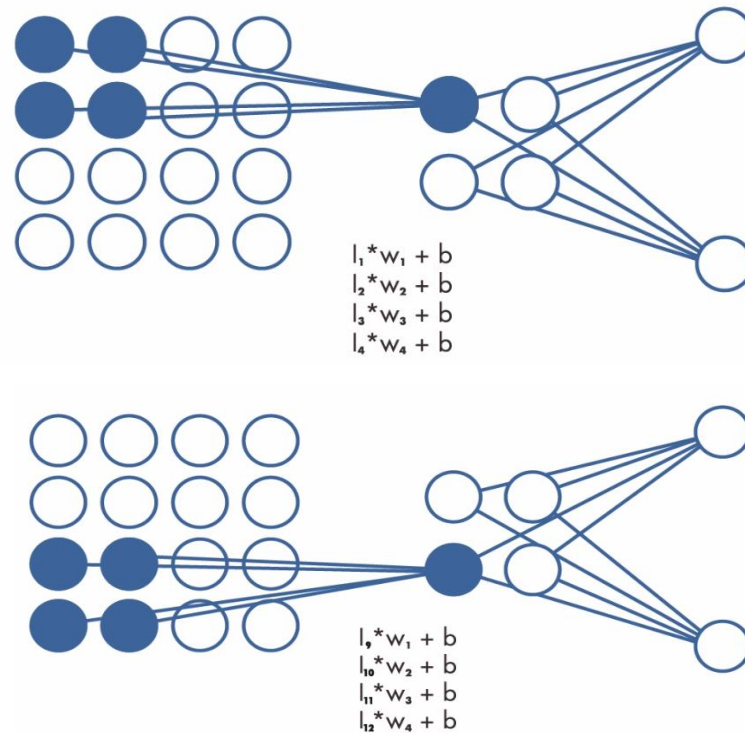


Fonte: adaptado de Mazza (2017)

Em uma determinada camada, os pesos e limiares, presentes no CNN, são compartilhados para todos os neurônios ocultos (Figura 4), o que faz que a mesma característica seja detectada, em diferentes regiões, por todos os neurônios ocultos (PATEL; PINGEL, 2017).

Para este trabalho, os valores dos pesos iniciais foram gerados aleatoriamente a partir uma distribuição gaussiana, com media zero e desvio padrão de 0,01 (padrão da toolbox Neural Network), assimilados durante o processo de treinamento e atualizados empregando-se a tecnica de retro-propagação (*backpropagation*).

Figura 4: Exemplo de compartilhamento de pesos(w_i) e limiares (b).

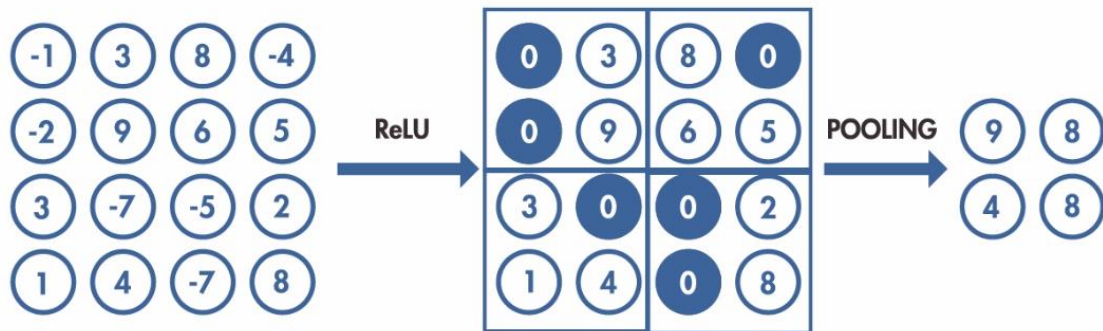


Fonte: Adaptado de (PATEL; PINGEL, 2017)

Ainda de acordo com Patel; Pingel (2017) outros dois conceitos das CNNs são ativação e pooling. Na ativação, as funções de ativação são funções utilizadas para que o modelo utilizado seja capaz de representar funções não lineares. Estas funções são aplicadas na saída de cada neurônio para gerar uma transformação. Uma função de ativação comumente utilizada é a Unidade Linear Rectificada, ou ReLU. Esta função mapeia as saídas dos neurônios para o maior valor positivo e caso estas sejam negativas, a saída é mapeada para zero.

O método de *pooling*, consiste em uma subamostragem não-linear dos dados, ou seja, a saída de uma pequena região de neurônios é condensada em uma única saída. Isto faz que as camadas seguintes sejam simplificadas e reduz o número de parâmetros necessários para o aprendizado e, conseqüentemente, reduz o tempo computacional das camadas superiores (PATEL; PINGEL, 2017). Estes métodos são ilustrados na Figura 5.

Figura 5: Exemplo de aplicação do ReLU e *pooling*.



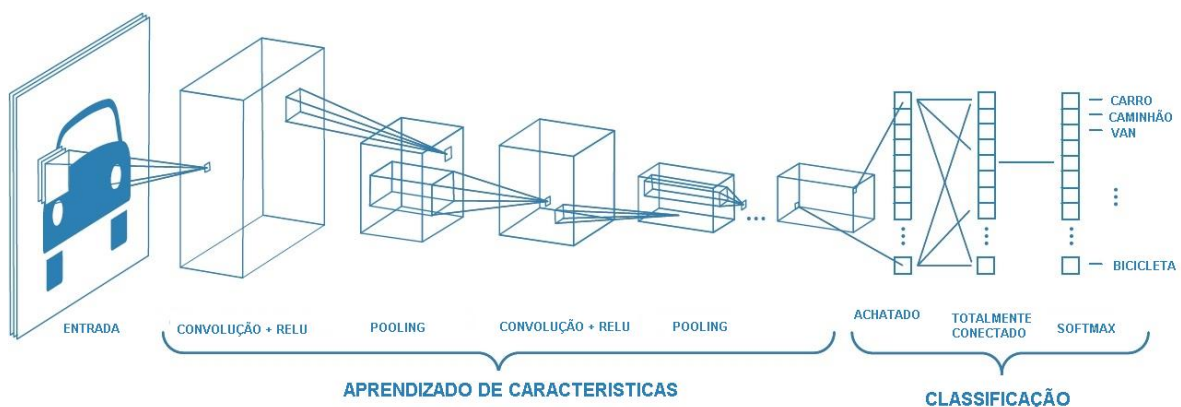
Fonte: (PATEL; PINGEL, 2017)

Em uma CNN, diferentes características podem ser detectadas por cada uma das camadas ocultas do modelo, podendo ela ter dezenas ou centenas de camadas. Com o aumento de camadas ocultas há também o aumento da complexidade dos recursos de aprendizado, na qual as últimas camadas podem aprender características mais complexas do que nas primeiras (PATEL; PINGEL, 2017).

As saídas da última camada oculta são conectadas a cada neurônio da camada totalmente conectada. Por fim, a esta última camada é aplicada a função *Softmax*, a qual normaliza as saídas de modo que o somatório de todas as saídas da rede possua valor 1, tendo como palavra reconhecida a que possui a maior probabilidade.

A CNN final pode ser representada pela Figura 6.

Figura 6: Diagrama de representação de uma rede neural convolucional.



Fonte: adaptado de (PATEL; PINGEL, 2017)

A ideia básica para utilização da CNN para reconhecimento de fala é usar um sistema no qual as entradas são partes locais da frequência. Estas entradas são aplicadas a filtros e, a partir de um conjunto deles, ocorre o aprendizado.

A saída é calculada deslocando-se cada um destes filtros, com seus pesos compartilhados ao longo da camada de entrada. O método de *pooling* é utilizado para diminuir a resolução das características de alto nível (ABDEL-HAMID; et al, 2013).

A vantagem de utilizar CNNs é seu alto poder de representatividade, pois pode representar precisamente novos exemplos, a partir de exemplos já observados. Estes modelos, porém, possuem grande número de parâmetros, logo estão sujeitos a sobre ajuste (*overfitting*) (MAZZA, 2017).

O *overfitting* pode ocorrer quando o modelo, ao invés de se ajustar a amostra verdadeira, se ajusta ao ruído ou quando decora as amostras observadas anteriormente, gerando uma grande diferença entre a acurácia do modelo de amostras já observadas e novas amostras.

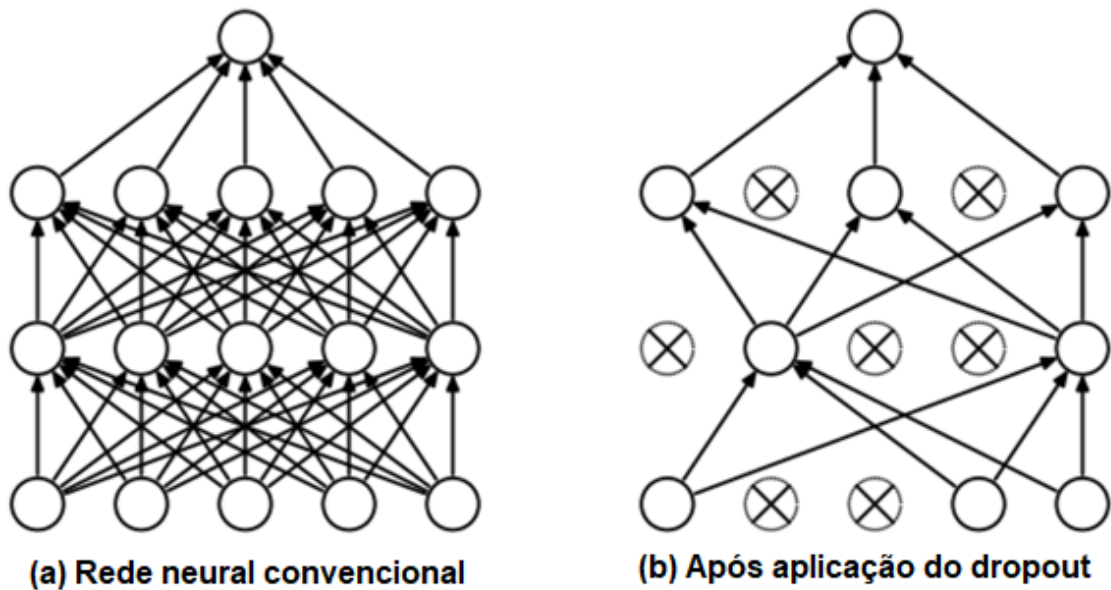
A fim de diminuir o *overfitting* e medir a capacidade real de generalização do modelo, utiliza-se a validação cruzada. Essa ferramenta separa a base de dados em três conjuntos: treino, validação e teste.

O conjunto de treinamento é utilizado para fazer o aprendizado do modelo e estimar os parâmetros internos. O conjunto de validação verifica o poder de generalização, ou seja, ela pode ser avaliada como uma aproximação da real acurácia do modelo. Por isso, a validação é utilizada para escolher os hiperparâmetros, ou seja, os parâmetros externos da rede.

Porém, a alteração dos hiperparâmetros na tentativa de diminuir um erro de validação, pode causar um viés, subestimando o erro de validação em relação ao erro real (CAWLEY, G. C e TALBOT, N. L., 2010). Com isso, utiliza-se o conjunto de teste para verificar se o modelo tem alta generalização, ou seja, verificar se taxa de erro dentro do conjunto de teste não seja superior ao de validação.

Métodos que reduzem o *overfitting* são chamados de regularização (KROGH, A. e HERTZ, J. A, 1992). Dentre os mais comuns podem ser citados o *dropout* (SRIVASTAVA et al, 2014) e o aumento de dados (*data augmentation*) (SALAMON, BELLO, 2017). O método *dropout* consiste na exclusão aleatória de unidades entre camadas de rede, dificultando a co-adaptação de parâmetros, a fim de prevenir que o modelo decore as entradas e se ajustem aos ruídos das amostras. Como pode ser visto na Figura 7.

Figura 7: Exemplo de aplicação de *dropout* em uma rede neural convencional



Fonte: (SRIVASTAVA et al, 2014)

O método de aumento de dados consiste em criar um banco de dados maior, modificando os exemplos já existentes por meio de uma ou mais deformações no conjunto de amostras (SALAMON, BELLO, 2017).

Aumenta-se artificialmente os dados com o objetivo de explorar as variações dos dados que podem ocorrer em casos reais. Este método é útil, pois uma maior quantidade de dados reduz as chances do modelo se adaptar a valores muito específicos dos exemplos de treino e perder a capacidade de generalização (Santos, 2017).

Uma das técnicas de aumento de dados que pode ser utilizada é a de mudança de tom (*pitch shift*), na qual aumenta ou diminui-se a frequência do áudio, gerando um áudio com o tom da voz modificado (SCHLÜTER, GRILL, 2015).

3. MATERIAIS E MÉTODOS

3.1 Materiais

Para o presente trabalho, o banco de dados constituiu-se por vozes cedidas pelo Grupo de Engenharia Médica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (GPEM/CNPq). Este banco é composto por frases ditas por diferentes pessoas que sofriam de dois tipos de patologia: nódulos vocais e edema de Reinke.

As gravações destas frases foram coletadas nas dependências do Ambulatório de Voz do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HC-FMUSP). Os registros da emissão vocal foram realizados em sala acusticamente tratada por fonoaudióloga com experiência em voz.

Para a captura, gravação e armazenamento dos sinais de voz foram utilizados o software Audacity em computador Pentium II acoplado a uma placa de som e um microfone headset unidirecional da marca AKG– C520 L. O microfone foi posicionado à distância de quatro centímetros da boca, formando com esta um ângulo de 45° à 90°. Os sinais acústicos foram armazenados em extensão de arquivo .WAV à taxa de 44100 amostras por segundos.

As palavras que compunham as frases deste banco de dados foram separadas manualmente, formando-se um outro banco composto por 20 palavras. Cada uma destas palavras continha vozes de 28 pessoas diferentes totalizando 560 amostras, nas quais foram divididas em dois grupos: treinamento e teste.

Utilizou-se 480 amostras para treinamento e 80 para testes. Para fazer a validação cruzada, dividiu-se então a base de treinamento (480 amostras) em 400 para treino e 80 para validação. As palavras utilizadas e o número de amostras para cada palavra podem ser observado no Quadro 2.

Quadro 2. Número de amostras de cada palavra

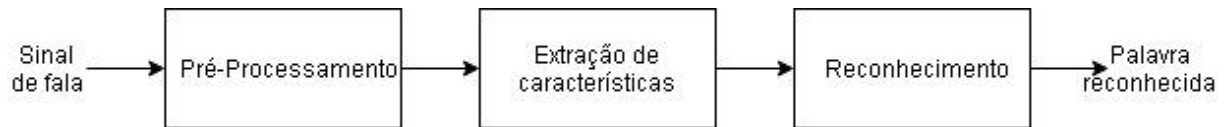
Palavra	Amostras de treinamento	Amostras de validação	Amostras de teste
Acabar	20	4	4
Agora	20	4	4
Anjo	20	4	4
Avião	20	4	4
Azul	20	4	4
Erika	20	4	4
Hora	20	4	4
Minha	20	4	4
Namorou	20	4	4
Olha	20	4	4
Papai	20	4	4
Pipoca	20	4	4
Quente	20	4	4
Sabe	20	4	4
Sambar	20	4	4
Sonia	20	4	4
Sozinha	20	4	4
Suco	20	4	4
Tomou	20	4	4
Trouxe	20	4	4
Total	400	80	80

Fonte: autoria própria.

3.2 Métodos

Os métodos utilizados podem ser representados por meio de um diagrama de blocos, conforme Figura 8.

Figura 8 – Diagrama de blocos de um sistema de reconhecimento de voz



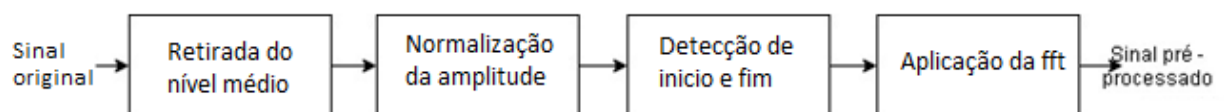
Fonte: Autoria própria

O sistema a ser utilizado é dividido em três etapas. A primeira etapa é a de pré-processamento, na qual as informações menos importantes do sinal serão eliminadas. A etapa seguinte é a de extração de características, que tem como objetivo diminuir o tamanho dos dados preservando apenas as características que melhor definem o sinal de fala. Por fim essas características passam pela fase de reconhecimento, gerando como saída a palavra reconhecida (ZANOTELLI, 2008).

3.2.1 Pré-processamento

Os dados adquiridos, necessários para o sistema de reconhecimento, sofrem perturbações do ambiente da gravação e do canal de comunicação. Por isso, é necessário que seja realizado um pré-processamento do sinal com o objetivo de filtrá-lo e deixá-lo mais próximo da fala pura (SILVA, 2009). Este procedimento é representado pelo diagrama da Figura 9.

Figura 9: Diagrama de blocos da fase de pré-processamento



Fonte: Autoria própria

Em grande parte das vezes o sinal de fala apresenta uma componente contínua que interfere na comparação em valores absolutas, isso gera a necessidade de retirar o nível médio, para que todas as amostras oscilem em torno do valor zero. Para essa retirada seja efetuada, calcula-se a média aritmética das amplitudes do sinal e depois subtrai-se cada amplitude desta média.

A normalização da amplitude tem relação com o volume do som. Esta etapa é efetuada para que o algoritmo de reconhecimento processe igualmente sons com

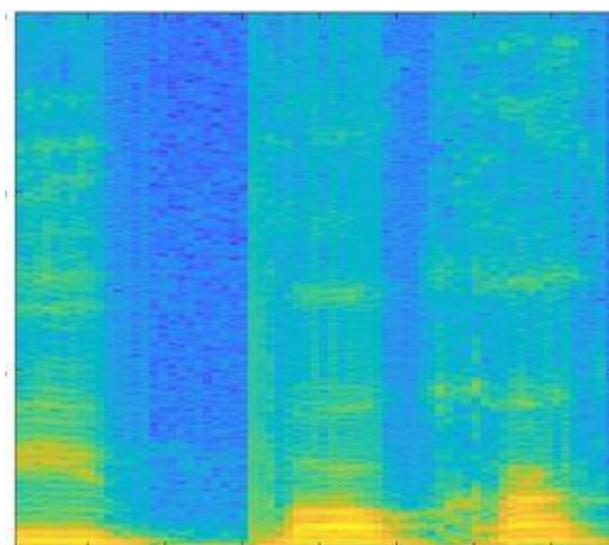
diferentes volumes, mais altos ou mais baixos. Para que isso seja realizado, é necessário dividir o valor de cada amostra do sinal pelo seu maior valor de amplitude, fazendo com que todos os valores de amplitude de todos os sinais fiquem na mesma faixa de valores e não tenham efeitos indesejáveis nos neurônios.

É feita, então, a detecção manual do início e fim da locução. Nesta parte, remove-se os períodos de silêncio existentes no início e no final do sinal. Esta detecção tem como finalidade diminuir a carga computacional e economizar tempo, além de remover possíveis ruídos ou sinais indesejados presentes no sinal da fala (SILVA, 2009); (SANTOS, 2014).

A fim de melhorar a capacidade de generalização do modelo e conseqüentemente diminuir o *overfitting*, empregou-se a técnica de mudança de tom (*pitch shift*), na qual aumentou-se em 20% a frequência da fala em todas as amostras de treinamento. Estes novos dados gerados foram adicionados aos anteriores, dobrando assim a base de dados de treinamento (SCHLÜTER, GRILL, 2015).

Para utilizar a entrada de áudio no modelo CNN proposto, aplicou-se uma transformada rápida de Fourier, utilizando-se a função *spectrogram* do Matlab. Esta função transforma a entrada de áudio em um espectrograma, na qual o eixo horizontal representa o tempo e o eixo vertical a frequência do áudio. Um exemplo do espectrograma criado pode ser visto na Figura 10.

Figura 10: Espectrograma da palavra pipoca criada pelo Matlab.



Fonte: autoria própria

Por fim, fez-se o redimensionamento do espectrograma para o tamanho de 64 pixels de altura, 64 pixels de largura e 3 canais de cores (RGB), com o objetivo de diminuir o esforço computacional.

3.3 Extração de características do sinal da fala

Para que o projeto de qualquer sistema de reconhecimento de fala tenha resultados representativos é de extrema relevância que ele contenha esta etapa de extração de informações. É por meio desta que será possível obter informações realmente significativas para a descrição do sinal de voz. Isso se deve ao fato de o sinal digital possuir uma grande quantidade de dados, como consequência, muitas das informações presentes no sinal podem ser redundantes ou não conter nenhuma importância para a distinção fonética (SANTOS, 2014).

A ideia básica desta etapa é computar o menor número possível de parâmetros que contenham apenas informações suficientes para caracterizar o sinal de fala e representar as unidades de fala com estes (SILVA, 2009).

Em uma CNN, as características do sinal de fala são extraídas aplicando-se filtros em pequenas partes da camada de entrada e deslocando-os ao longo de toda a camada, criando assim, um mapa com as características desta.

Atualmente, não existe na literatura uma maneira de determinar a melhor topologia de uma CNN para todos os problemas, pois os resultados dependem de cada modelo em específico, logo a escolha da topologia de uma CNN é um processo empírico. Portanto, foram testadas diversas topologias para encontrar a que obtivesse uma maior acurácia.

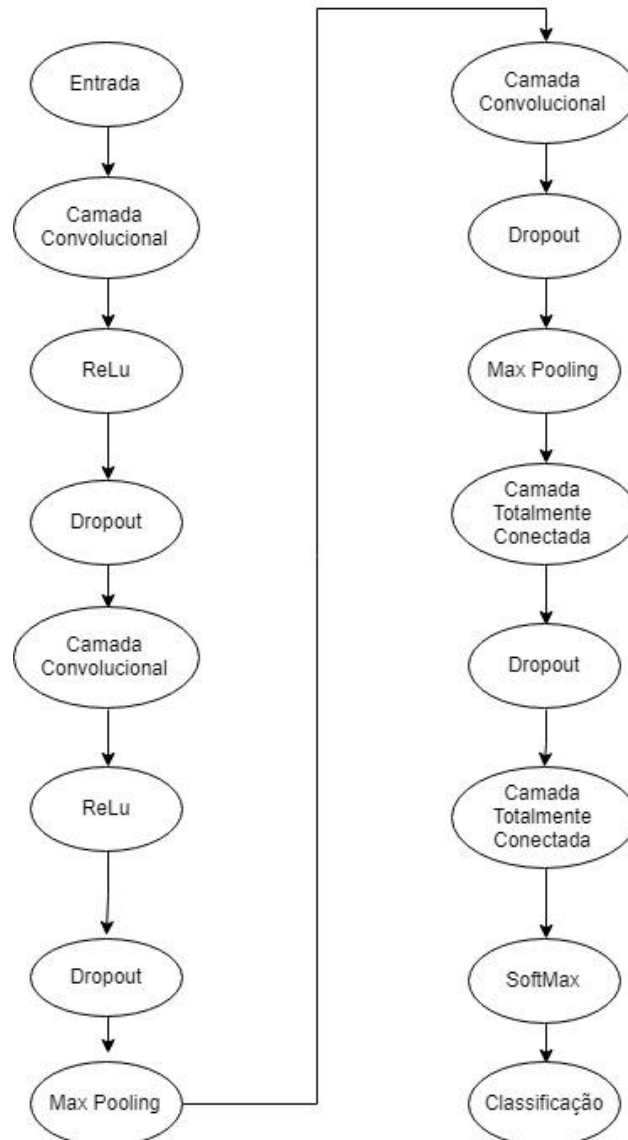
Durante a configuração das CNNs, baseou-se na topologia descrita em Miyazaki (2017) e definiu-se como fixo, os seguintes Hiperparâmetros:

- Passo da camada convolucional = 1;
- Passo da camada de *Max Pooling* = 2;
- Tamanho do filtro do *Max Pooling* = 2;
- Número de épocas = 180;
- Dropout = 0,5;
- Taxa de aprendizado = 0,0001;

- Número de mapas de características.
 - Para duas camadas convolucionais:
 - 1ª camada: 32 e 2ª camada: 64
 - Para três camadas convolucionais:
 - 1ª camada: 16, 2ª camada: 32 e 3ª camada: 64

Para selecionar o modelo de CNN deste trabalho, variou-se o número de camadas convolucionais, de camadas totalmente conectadas e o tamanho do filtro. Sendo a arquitetura mais completa a de três camadas convolucionais e duas totalmente conectadas como pode ser observada na Figura 11. As outras arquiteturas descritas na Tabela 3 são pequenas variações desta.

Figura 11 - Arquitetura completa para o modelo CNN proposto.



Fonte: autoria própria

3.4 Reconhecimento

Neste trabalho o reconhecimento será realizado utilizando técnicas de classificação de padrões. Assim para a realização desta etapa o sistema será dividido em duas fases: treinamento e teste.

De acordo com Silva; Spatti; Flauzino (2010) e Russell; Norvig (1995) a fase de treinamento consiste na atualização dos pesos sinápticos e limiares com o propósito de aproximar o vetor de saída da saída esperada. Este processo ocorre em duas etapas principais: a propagação adiante (*forward-propagation*) e retro-propagação (*backpropagation*).

Na etapa de propagação insere-se os sinais de uma amostra de treinamento na entrada da rede e a partir dos filtros estipulados, calcula-se os mapas de características da camada de entrada e das camadas ocultas. No final deste processo, utiliza-se a última camada para gerar uma saída com os valores estimados pela rede.

Após a etapa de propagação, determina-se o desempenho da rede comparando a distância que os valores de saída estão dos desejados. Este processo é executado calculando-se cada Entropia Cruzada (*Cross Entropy*) e fazendo a média deles de toda a rede.

O método *Cross Entropy* é descrito pela Equação (1) na qual, n é o número de classes a serem reconhecidas, S é o vetor de saída da rede e L o vetor que representa cada uma das classes esperadas (representadas por vetores binários).

$$D(S, L) = - \sum_i^n L_i \text{Log}(S_i) \quad (1)$$

Caso este desempenho não seja suficiente, inicia-se a fase de retro-propagação, onde se deseja minimizar o erro de estimação. Esta fase utiliza o método gradiente descendente estocástico para fazer a minimização, modificando os pesos dos filtros, de acordo com a taxa de aprendizado da rede, com o objetivo de produzir a maior queda ao longo da superfície de erro, continuando até encontrar um erro mínimo local (MIYAZAKI, 2017).

Terminado a fase de treinamento, ocorre a fase de teste, na qual acrescentam-se novos dados de entrada e a rede deverá ser capaz de generalizar os exemplos apresentados e classificá-los de acordo com classes previamente estabelecidas,

fazendo assim o reconhecimento das palavras (DA SILVA; SPATTI; FLAUZINO, 2010).

4. RESULTADOS E DISCUSSÕES

Neste capítulo, serão apresentados os principais resultados obtidos com o modelo de reconhecimento de fala. Expõem-se os dados de validação, com as porcentagens de precisão de reconhecimento, e avaliam-se o desempenho do modelo com a maior acurácia.

Durante a fase de reconhecimento dos modelos CNNs apresentados, testou-se 5 vezes cada uma das topologias, obtendo-se como resultado o Quadro 3, que apresenta a maior acurácia obtida de cada uma das amostras para cada configuração da rede.

Quadro 3: Acurácia das amostras de treinamento, validação e teste para cada configuração da CNN.

Número de camadas convolucionais	Número de camadas totalmente conectadas	Tamanho do filtro	Acurácia Treinamento	Acurácia Validação	Acurácia Teste
2	1	6x6	0,4688	0,8000	0,7750
	1	8x8	0,4531	0,8125	0,7000
	2	6x6	0,8281	0,8125	0,8000
	2	8x8	0,8359	0,8125	0,7750
3	1	6x6	0,6016	0,8125	0,7750
	1	8x8	0,5625	0,7875	0,7875
	2	6x6	0,7891	0,8250	0,8250
	2	8x8	0,8438	0,8125	0,7875

Fonte: autoria própria

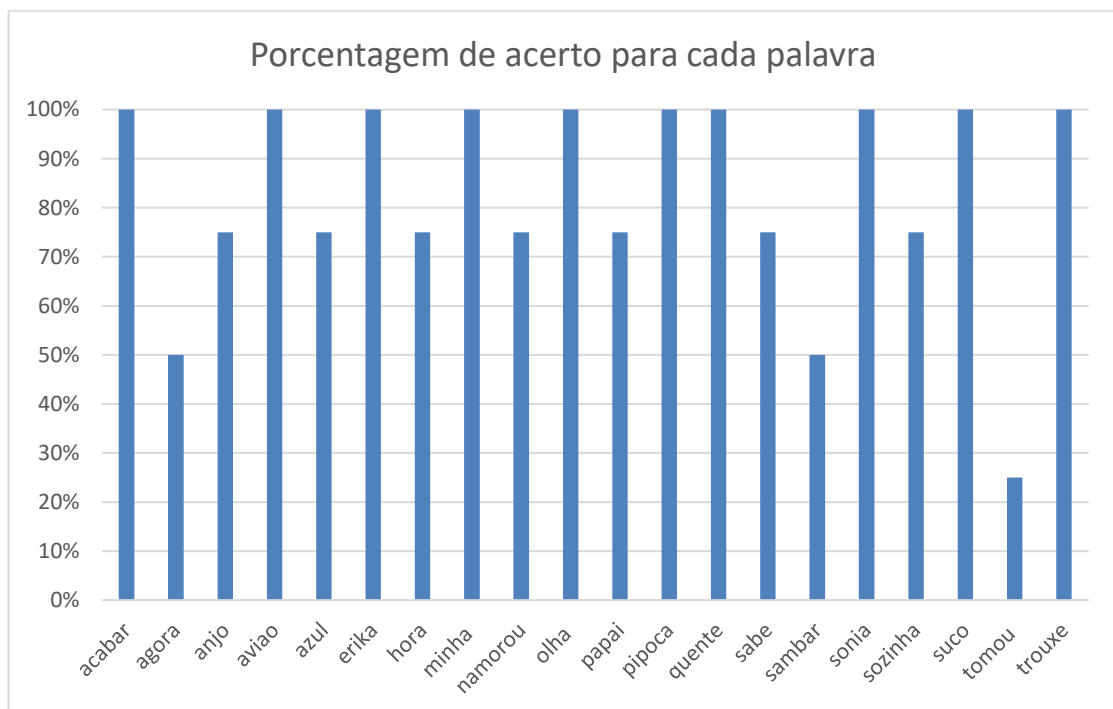
Analisando as variáveis, separadamente, do Quadro 3, pode-se notar que as topologias com 3 camadas convolucionais e 2 camadas totalmente conectadas apresentaram melhores resultados no conjunto de treino. Sendo a topologia com 3 camadas convolucionais, 2 camadas totalmente conectadas e filtro de tamanho 6x6 a que obteve o melhor resultado, com 82,50% de acurácia.

Estes resultados eram esperados, pois, topologias com mais camadas convolucionais podem representar características mais complexas e de acordo com Lecun (1998), apesar de uma única camada totalmente conectada ser suficiente, em situações práticas, duas destas camadas produzem melhores resultados.

O tamanho do filtro tem relação com a localização das características chave dos dados de entrada. Um filtro de tamanho maior pode ignorar detalhes essenciais, enquanto um filtro de tamanho menor poderia fornecer mais informações, levando a mais confusão. Para este problema, o filtro de tamanho 6x6 demonstrou ser mais eficiente para encontrar as características principais, enquanto o filtro 8x8 perdeu algumas das características essenciais para o reconhecimento e teve uma acurácia menor.

Para avaliar o melhor resultado, na qual a acurácia geral foi de 82,50%, verificou-se a taxa de reconhecimento apresentada na Figura 12, na qual as linhas azuis representam as palavras reconhecidas corretamente. Nela pode-se perceber que 10 palavras apresentaram percentagens de acerto de 100%, 7 de 75%, 2 de 50% e uma de 25%.

Figura 12 – Porcentagem de acerto para cada palavra



Fonte: autoria própria

O desempenho (erros e acertos) da topologia treinada são apresentados na matriz de confusão, visualizada no Quatro 4. Esta matriz apresenta informações da classificação efetuada (KOHAVI; PROVOST, 1998). É formada por linhas que contém as classes previstas e colunas que representam qual das classes foi classificada. Já

os números presentes na matriz apresentam a quantidade de amostras que foram classificadas para aquela classe.

A diagonal principal da matriz identifica o número de imagens corretamente classificadas e os restantes elementos representam imagens classificadas incorretamente.

A hipótese de o modelo ter errado 17,50%, ou seja, 14 das 80 amostras de teste, está associada à qualidade e quantidade das amostras de entrada. Como os espectrogramas foram reduzidos para diminuir a complexidade computacional, foi perdida uma parte da resolução da imagem e conseqüentemente informações da voz.

Com relação a quantidade das amostras, a baixa taxa de acertos, se deve à limitação no tamanho do banco de dados utilizado. Isso ocorre pois a CNN necessita de uma grande quantidade de amostras para conseguir obter um resultado satisfatório.

Quadro 4: Matriz de confusão do resultado com maior acurácia

acabar	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
agora	1	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Anjo	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
avião	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Azul	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Erika	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hora	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
minha	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
namorou	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0
olha	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
papai	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0
pipoca	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
quente	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0
sabe	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0
sambar	0	1	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0
Sonia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
sozinha	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
suco	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
tomou	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0
trouxe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	acabar	agora	Anjo	avião	Azul	Erika	hora	minha	namorou	olha	papai	pipoca	quente	sabe	sambar	Sonia	sozinha	suco	tomou	trouxe
TOTAL	5	4	5	4	5	5	4	4	6	4	3	4	4	4	2	4	4	4	2	4

Fonte: autoria própria

5. CONSIDERAÇÕES FINAIS

Atualmente foram desenvolvidas diversas aplicações utilizando-se reconhecimento de fala, porém pessoas com algum tipo de disfonia (rouquidão) na voz não conseguem utilizar essa tecnologia em sua totalidade. Logo, este trabalho possui o intuito de avaliar um modelo que pudesse tratar as distorções presentes na voz destas pessoas.

Para alcançar este objetivo, utilizou-se uma rede neural convolucional pois, além de apresentar resultados robustos no reconhecimento de fala, foram desenvolvidos para serem invariantes a deslocamento, escala e distorção.

No capítulo de materiais e métodos foi apresentado o banco de vozes deste trabalho, no qual consiste em 20 palavras ditas por 28 pessoas diferentes, totalizando 560 amostras, sendo elas separadas em 20 amostras para treino, 4 para validação e 4 para teste. Nestas 20 amostras de treino, aplicou-se aumento de dados, dobrando o conjunto. Foi então aplicado uma transformada rápida de Fourier nos áudios para gerar o espectrograma que foi usado para alimentar os modelos.

Foram comparadas topologias diferentes de CNN a fim de se selecionar o mais apropriado para este problema. Por fim, o modelo na topologia com 3 camadas convolucionais, 2 camadas totalmente conectadas e tamanho de filtro igual a 6x6, obteve o melhor resultado. Esta CNN apresentou uma precisão geral de (82,50%), mostrando-se eficaz para o reconhecimento de palavras de pessoas com rouquidão, para vocabulário pequeno, mesmo com um baixo volume de amostras.

5.1 Trabalhos Futuros

Este trabalho tinha como proposta analisar e fazer o reconhecimento de palavras faladas por pessoas com algum tipo de rouquidão. Apesar de o modelo CNN apresentar desempenho satisfatório, não apresentou seu desempenho máximo. Isso se deve a alguns fatores como:

- Não foram abordados todos os métodos para redução de *overfitting*.
- Baixa quantidade de dados para treinamento.
- Otimização manual do modelo, o que pode não ter abordado todos os resultados possíveis.

Logo, sugere-se para trabalhos futuros:

- Analisar outras técnicas de redução de overfitting.
- Ampliar o conjunto de dados para treinamento e validação, com o intuito de obter novas medições de precisão e desempenho da topologia proposta.
- Abordar outras topologias de CNN.

REFERÊNCIAS

- ABDEL-HAMID, O. et al. **Deep segmental neural networks for speech recognition**. In: Interspeech. 2013. p. 70.
- BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York, Springer, 2006
- CAWLEY, G. C., & Talbot, N. L. **On over-fitting in model selection and subsequent selection bias in performance evaluation**. Journal of Machine Learning Research, 11(Jul), p. 2079-2107, 2010.
- COLE, R. M., et al. (Eds.). **Survey of the State of the Art in Human Language Technology**. 1996. Publicação eletrônica. Disponível em: <<http://www.cslu.ogi.edu/HLTsurvey/HLTsurvey.html>>. Acesso em 26 de maio de 2017.
- DAHL, G. E., et al. **Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition**. IEEE Transactions on Audio, Speech, and Language Processing 20.1, 30-42, 2012.
- DA SILVA, I.N., SPATTI D. H.; FLAUZINO, R.. **Redes neurais artificiais para engenharia e ciências aplicadas: curso prático**, Artliber Editora Ltda, São Paulo, SP, Brasil. 2010
- DENG, Li.; DONG, Yu. **Deep learning: methods and applications**. Foundations and Trends in Signal Processing 7.3–4, 197-387, 2014.
- GARG, A., SHARMA, P. **Survey on acoustic modeling and feature extraction for speech recognition**. In Computing for Sustainable Global Development (INDIACom), 3rd International Conference on(pp. 2291-2295). IEEE. 2016.
- GOODFELLOW, I., BENGIO, Y., COURVILLE, A. **Deep learning, 2016**. MIT Press.[Online]. Disponível em: <<http://www.deeplearningbook.org/>>. Acessado em 26 de maio de 2017.
- HAYKIN, S. S. **Neural Networks: A Comprehensive Foundation**. New Jersey, US: Prentice Hall, 1999.
- HINTON, G., et al. **Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups**. IEEE Signal Processing Magazine 29.6, 82-97, 2012.
- INCE, A. N., ed. **Digital Speech Processing: Speech Coding, Synthesis and Recognition**. Vol. 155. Springer Science & Business Media, 2013.
- KOHAVI, R.; PROVOST, F. **Glossary of Terms. Machine Learning**, v. 30, n. (2-3), p. 271– 274, 1998.

KROGH, A.; HERTZ, J. A. **A simple weight decay can improve generalization.** In: Advances in neural information processing systems. 1992. p. 950-957.

LECUN, Y.; BENGIO, Y.; HINTON, G., **Deep learning.** nature, v. 521, n. 7553, p. 436, 2015.

LECUN, Y. et al. **Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks,** v. 3361, n. 10, p. 1995, 1995.

LECUN, Y. et al. **Gradient-based learning applied to document recognition.** Proceedings of the IEEE, v. 86, n. 11, p. 2278-2324, 1998.

LI, B. **Building pattern classifiers using convolutional neural networks.** In: Neural Networks. IJCNN '99. International Joint Conference on. [S.l.: s.n.]. p. 3081–3085, 1999.

MAMEDE, N. J., et al., **Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, Proceedings, (Vol. 2721).** Springer, 2003.

MARSLAND, S. **Machine learning: an algorithmic perspective.** 2st. Ed. CRC press, 2015

MAZZA, L. O. **Aplicação de Redes Neurais Convolucionais Densamente Conectadas no Processamento digital de imagens para remoção de ruído Gaussiano.** TCC (Graduação) – Curso de engenharia de Controle e Automação, Escola Politécnica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2017.

MEHRYAR, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of Machine Learning,** MIT Press. 2012

MIYAZAKI, C. K. **Redes neurais convolucionais para aprendizagem e reconhecimento de objetos 3d.** 79 f. TCC (Graduação) - Curso de Engenharia Elétrica com Ênfase em Sistemas de Energia e Automação, Escola de Engenharia de São Carlos da Universidade de São Paulo, São Carlos, 2017.

MITCHELL, T. M. **Machine learning.** Burr Ridge, IL: McGraw Hill, 1997.

MITRA, S. K.; KUO, Y. **Digital signal processing: a computer-based approach.** Vol. 2. New York, US: McGraw-Hill, 2006

MOHAMED, A., DAHL, G. E., HINTON, G.. **Acoustic modeling using deep belief networks.** IEEE Transactions on Audio, Speech, and Language Processing 20.1 14-22, 2012.

PATEL, S. **Introduction to Deep Learning: What is Deep Learning?.** Disponível em: <<https://www.mathworks.com/videos/introduction-to-deep-learning-machine-learning-vs-deep-learning-1489503513018.html>> Acesso em 23 de maio de 2017

PATEL, S., PINGEL, J. **Introduction to Deep Learning: What Are Convolutional Neural Networks?**. Disponível em: <<https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html>> Acesso em 23 de maio de 2017

PINGEL, J. **Introduction to Deep Learning: Machine Learning vs. Deep Learning**. Disponível em: <<https://www.mathworks.com/videos/introduction-to-deep-learning-machine-learning-vs-deep-learning-1489503513018.html>> Acesso em 23 de maio de 2017.

PROAKIS, J. G.; MANOLAKIS, D. G. **Digital Signal Processing principle, Algorithms, and Applications**, Printice-Hall. Inc, New Jersey, 1996.

RUSSELL, S. J., NORVIG, P. **Artificial Intelligence: A modern approach**. Prentice-Hall, Egnlewood Cliffs, 1995.

SALAMON, J., BELLO, J. P. **Deep convolutional neural networks and data augmentation for environmental sound classification**. IEEE Signal Processing Letters, v. 24, n. 3, p. 279-283, 2017.

SANTOS, C. M. M. dos. **Desenvolvimento de um sistema de reconhecimento de fala usando modelos ocultos de Markov**. 2014. 54 f. TCC (Graduação) - Curso de Engenharia Elétrica, Universidade Tecnológica Federal do Paraná, Cornélio Procopio, 2014.

SANTOS, C. A. S dos. **Reconhecimento de imagens de marcas de gado utilizando redes neurais convolucionais e máquinas de vetores de suporte**. Dissertação, Engenharia Elétrica da Universidade Federal do Pampa, Alegrete, 2017.

SCHLÜTER, J., GRILL, T. **Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks**. In: ISMIR. 2015. p. 121-126.

SEIDE, F., Li, G., Yu, D. **Conversational Speech Transcription Using Context-Dependent Deep Neural Networks**. In: Twelfth Annual Conference of the International Speech Communication Association. 2011.

SILVA, A. G. da. **Reconhecimento de voz para palavras isoladas**. Monografia (Engenharia da Computação) — Universidade Federal de Pernambuco, Recife, 2009.

SRIVASTAVA, N. et al. Dropout: **A simple way to prevent neural networks from overfitting**. The Journal of Machine Learning Research, v. 15, n. 1, p. 1929-1958, 2014.

STEPP, R. E., RYSZARD S. M. **Conceptual clustering: Inventing goal-oriented classifications of structured objects**. Machine learning: An artificial intelligence approach 2, 1986.

STUCKLESS, R. **Developments in real-time speech-to-text communication for people with impaired hearing**. In M. Ross(Ed.), Communication access for people with hearing loss (pp.197-226). Baltimore, MD: York Press. 1994.

VALIATI, J. F. **Reconhecimento de voz para comandos de direcionamento por meio de redes neurais**. Dissertação (Mestrado) —Universidade Federal do Rio Grande do Sul, 2000.

VACHER, M. et al. **Acquisition et reconnaissance automatique d'expressions et d'appels vocaux dans un habitat**. In: JEP-TALN-RECITAL 2016. 2016. p. 28-36.

ZANOTELLI, T. **Reconhecimento de fala de locutor restrito para acionamento de dispositivos usando Modelos Ocultos de Markov**. 85 p. Monografia (Engenharia Elétrica) — Universidade Federal de Viçosa, Viçosa, 2008.