UNIVERSIDADE FEDERAL DO ABC

CECS - CENTRO DE ENGENHARIA, MODELAGEM E CIÊNCIAS SOCIAIS APLICADAS

CURSO DE GRADUAÇÃO EM ENGENHARIA DE INFORMAÇÃO

DETECÇÃO DE ATIVIDADE DE VOZ EM ÁUDIOS DE BAIXA QUALIDADE

VICTOR COSTA BERALDO

SANTO ANDRÉ

2018

UNIVERSIDADE FEDERAL DO ABC

CURSO DE GRADUAÇÃO EM ENGENHARIA DE INFORMAÇÃO

DETECÇÃO DE ATIVIDADE DE VOZ EM ÁUDIOS DE BAIXA QUALIDADE

VICTOR COSTA BERALDO

SANTO ANDRÉ

CECS - CENTRO DE ENGENHARIA, MODELAGEM E CIÊNCIAS SOCIAIS APLICADAS

2018

DETECÇÃO DE ATIVIDADE DE VOZ EM ÁUDIOS DE BAIXA QUALIDADE

Trabalho apresentado como requisito parcial para a Conclusão do Curso de Engenharia de Informação da Universidade Federal do ABC.

Área de conhecimento: Engenharia de Informação

Orientador: Prof. Dr. Murilo Bellezoni Loiola

SANTO ANDRÉ

CECS - CENTRO DE ENGENHARIA, MODELAGEM E CIÊNCIAS SOCIAIS APLICADAS

2018

RESUMO

A detecção de atividade de voz em áudios apresenta-se extremamente importante na construção de modelos de processamento de voz como reconhecimento de falantes e de fala. Atualmente, essa tarefa é feita utilizando limiares de energia do sinal de voz e sem a utilização de algoritmos de aprendizado de máquina para obtenção de um detector mais robusto, principalmente em ambientes ruidosos e com áudios gravados com baixa qualidade.

Este trabalho, portanto, apresenta uma metodologia para uma melhor detecção de atividade de voz utilizando modelos baseados em árvores de decisão como RF (*Random Forest*) e GB (*Grandient Boosting*) com atributos extraídos utilizando MFCCs (*Mel Frequency Cepstral Coefficients*). Os modelos foram treinados utilizando uma base de dados pública utilizada no trabalho [Kim 2017].

Palavras-chave: Detecção de Atividade de Voz; Processamento de Sinais; Aprendizado de Máquina; Sistemas Inteligentes.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

VAD	Detecção de Atividade de Voz (*Voice Activity Detection*)
RF	*Random Forest*
GB	*Gradient Boosting*
ML	*Machine Learning*
TDF	Transformada Discreta de Fourrier
DCT	Transformada Discreta de Cosseno (*Discrete Cosine Transform*)
MFCC	*Mel Frequency Cepstral Coefficients*

SUMÁRIO

1 INTRODUÇÃO

A voz é a portadora natal mais eficiente de informação do ser humano e, provavelmente, se tornará o próximo principal modo de interação humano-computador [Wang, Xu & Li 2011]. Existem diversas aplicações utilizando processamento de voz para intensificar essa interação, e para isso é preciso detectar a presença de voz em um sinal de áudio.

O detector de atividade de voz (VAD, do inglês Voice-Activity-Detection) é um algoritmo que tem como objetivo separar um sinal de áudio em trechos contendo voz e trechos sem voz. Ele serve como uma ferramenta importante para a análise de reconhecimento do falante, afetando diretamente o desempenho do modelo a ser implementado [Wang, Xu & Li 2011].

O VAD faz parte do estágio de pré-processamento para os principais aplicações em processamento de voz. As aplicações se estendem às comunicações móveis, transmissão de sinais de voz pela internet e supressão de ruídos em aparelhos auditivos digitais [Elton, Vasuki & Mohanalin 2016].

Este trabalho propõe uma nova técnica de VAD com o objetivo de detectar regiões com atividade de voz, a partir de um sinal de áudio com taxa de amostragem de 8kHz, características também vistas em áudios utilizado em sistemas de telefonia [Daengsi *et al.* 2012]. Para este propósito, serão utilizadas técnicas de aprendizado de máquina como algoritmos baseados em árvores de decisão: RF (*Random Forest*) e GB (*Gradient Boosting*) utilizando como variáveis informações extraídas dos MFCCs (*Mel-Frequency Cepstral Coefficients*) utilizando o pacote python speech features [Python Speech Features 2018], feito na linguagem de programação Python.

O estudo feito neste projeto utiliza-se de tecnologias como o aprendizado de máquina e suas aplicações com dados de voz, que tendem a ser incorporadas cada vez mais na indústria, como ocorre em setores que utilizam ligações telefônicas para o atendimento ao cliente. A Figura 1 mostra tecnologias emergentes em um gráfico que relaciona a expectativa de uma tecnologia ser incorporada na indústria, e em quanto tempo aproximadamente esta determinada tecnologia vai demorar para passar pelo intervalo onde as expectativas estão infladas, como é o caso do Aprendizado de Máquina (Machine Learning), para um intervalo onde a tecnologia é plenamente adotada no mercado. Esse tempo segundo o "Hype Cylcle" pode demorar aproximadamente de 2 a 5 anos, a partir de 2017. Logo essa tecnologia tende a se instaurar a partir do próximo ano (2019) de maneira ainda mais presente no mercado [Gartner 2017].

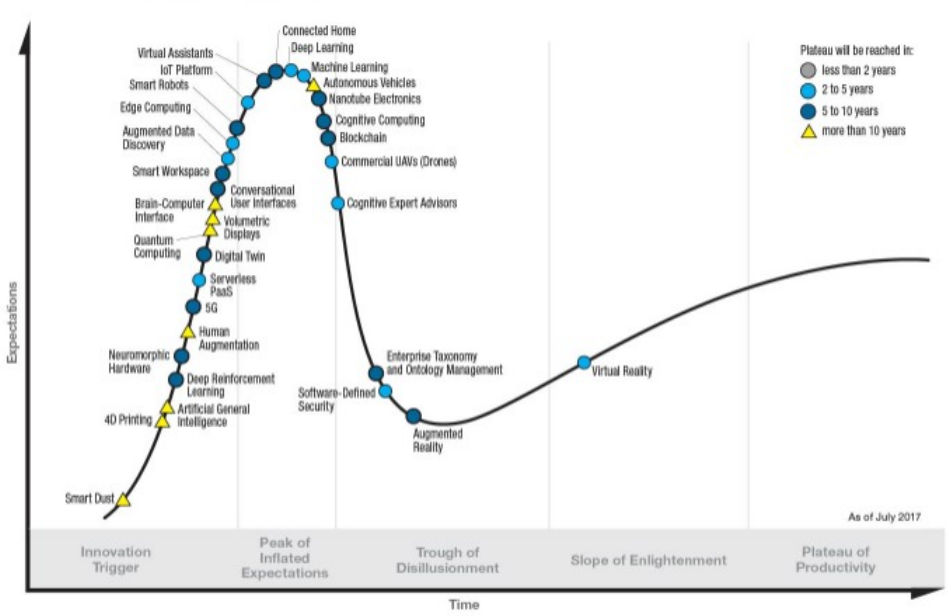


Figura 1 – Gráfico da expectativa de tecnologias emergentes em relação ao tempo “Hype Cycle”

Fonte: [Gartner 2017]

1.1 OBJETIVOS

Os principais objetivos deste projeto são:

- Aplicação da técnica de detecção de voz a partir de um modelo de classificação baseado em árvores de decisão em áudios com 8kHz e baixa SNR;
- Comparar os resultados com outros modelos de detecção de voz;
- Mostrar as vantagens e desvantagens da criação de um modelo de aprendizado de máquina com poucos dados rotulados, mostrando a necessidade de uma massa maior de dados para obter maior generalização.

1.2 ORGANIZAÇÃO DO TRABALHO

O estudo realizado neste trabalho foi feito começando pelo Capítulo 2, onde é exporto uma revisão de literatura, com objetivo de dar sustentação para o melhor entendimento do tema, tornando viável o entendimento do método abordado. O Capítulo 3 traz a metodologia e as técnicas utilizadas neste projeto. Especificações da base de dados e arquitetura do sistema também são explicadas, deixando claro a maneira com que o estudo foi feito e como foram extraídos os resultados. Os resultados, por sua vez que são explorados no Capítulo 4, com a apresentação dos resultados experimentais e a discussão sobre os mesmos, incluindo a comparação dos métodos propostos com outros modelos já

treinados. O trabalho é finalizado com o Capítulo 5, trazendo a conclusão e insumos para obtenção de melhores resultados em trabalhos futuros.

2 REVISÃO DA LITERATURA

2.1 A ANATOMIA DA FALA E DA AUDIÇÃO

A seção corrente aborda de maneira simplificada o modo como o ser humano produz a fala e como ouve, com base na referência [Beigi 2011]. Uma vez que esse mecanismo é melhor entendido, torna-se possível a criação de sistemas artificiais que possam distinguir características e realizar tarefas como detecção de voz, reconhecimento de falante e reconhecimento de fala.

2.1.1 Aparelho Fonador Humano

A Figura 2 mostra uma secsão sagital do nariz, boca, faringe e laringe. Essa é a porção superior e a mais significante parte da produção da fala dos seres humanos. Os pulmões são os únicos faltantes, uma vez que a principal função é fornecer a diferença de pressão necessária para produzir a fala.

Começando a análise da Figura 2 na região inferior, é possível ver uma seção da traqueia acompanhada pela seção da laringe, que inclui as cordas vocais (cordas vocais), acompanhadas de cartilagem dos dois lados, controladas por músculos que podem abrí-las e fecha-las totalmente.

A abertura das cordas vocais tem formato triangular com uma sensível inclinação. A Figura 3 mostra esta abertura com maiores detalhes. A área imediatamente abaixo das cordas vocais é chamada de glote, é o ponto de partida do controle para articulação. A glote é transição entre a laringe e a faringe. Quanto mais o formato se assemelha a um "v"na abertura das cordas vocais, mais tensos os músculos estão. Quando aberta completamente, os músculos estão completamente relaxados e proporcionam maior resistência ao fluxo de ar da traqueia para a parte inferior da laringe e faringe. Dependendo da tensão nos músculos das cordas vocais e a diferença na pressão do ar entre a traqueia e a faringe, tanto sopros de ar, quanto sons sonoros podem ser produzidos.

A faringe começa com seções de formas irregulares, após o ar passar pelas cordas vocais e quase imediatamente excitando a laringe, dando origem ao chamado trato vocal. Todos os controles de articulação começam aqui. Dependendo do modo com que o ar passa pelas cordas vocais, pode passar por diferentes seções da faringe, produzindo diferentes tipos de sons .

O próximo ponto interessante, à medida que o ar se desloca para cima no trato vocal é a epiglote, levando o ar do dorso até a língua, músculo que define um limite móvel para parte oral da faringe. Na parte oral da faringe, o ar pode passar por duas cavidades

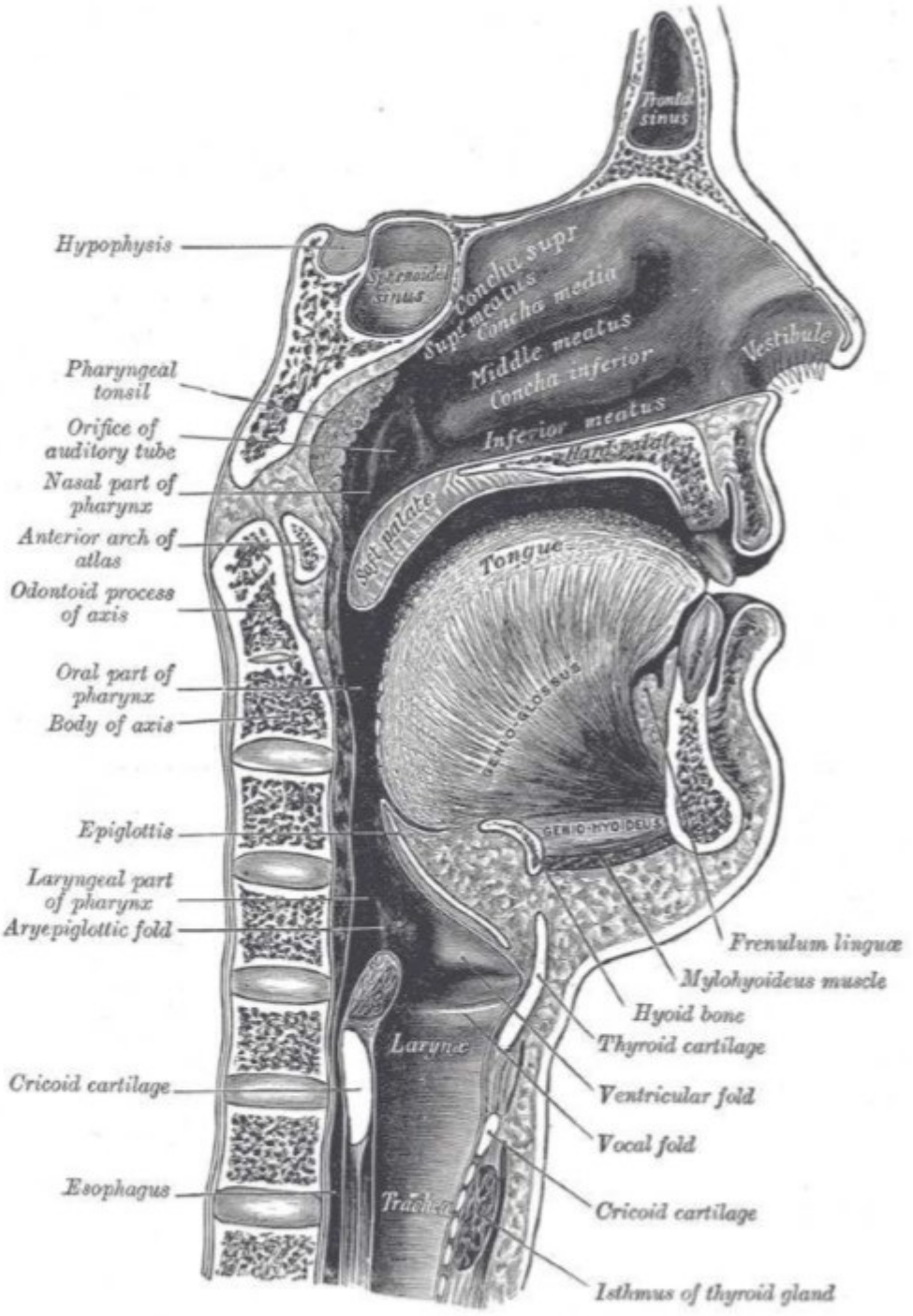


Figura 2 – Sessão sagital do nariz, boca faringe e laringe.

Fonte: [Gray & Lewis 1918]

diferentes, a cavidade oral e a nasal, sendo chamadas parte oral da faringe e parte nasal da faringe .

A úvula e o palato mole (véu do palato) são os principais responsáveis pelo desvio do ar para cavidade nasal ou oral. A cavidade nasal é essencialmente uma caixa de som que dissipa principalmente energia na fala, liberando o ar para pressão ambiente através

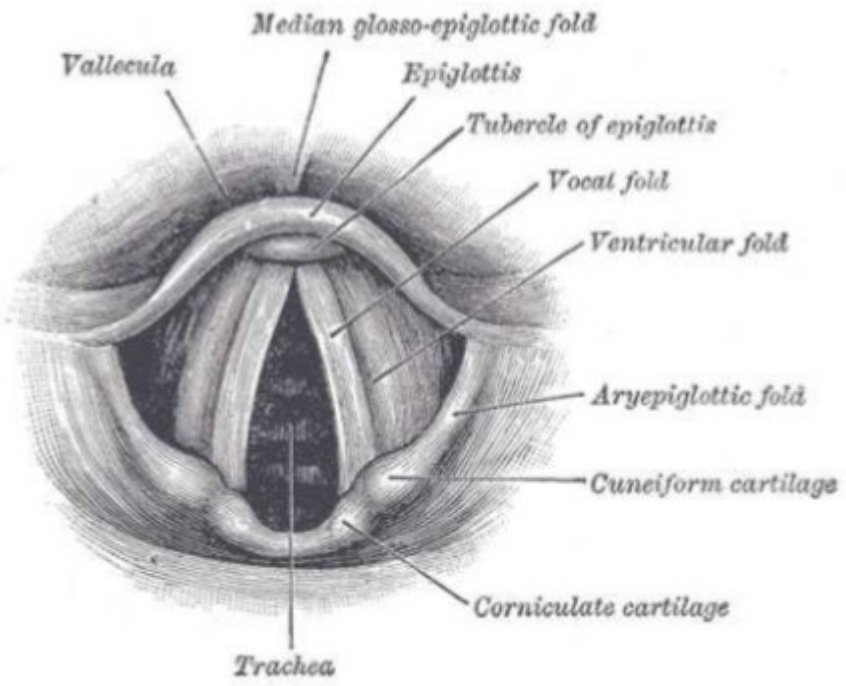


Figura 3 – Interior da laringe, visão Laringoscópica das cordas vocais.

Fonte: [Gray & Lewis 1918]

do vestíbulo .

Caso a passagem do ar para a cavidade nasal seja bloqueada usando a forma da língua e a posição da úvula e do véu do palato, o fluxo de ar passará ao longo da superfície da língua para a cavidade oral, restringida pelo topo do palato duro, que estende até os dentes superiores .

O ponto de saída do ar pode ser a cavidade oral ou nasal de saída. O ar já passou pelo trato vocal até a cavidade oral, neste ponto encontra obstáculos, que podem moldar a articulação: os dentes, lábios e também a língua.

2.1.2 Sistema Auditivo Humano

O sistema auditivo inclui um sistema mecânico e um sistema nervoso. Nesta seção será abordada, predominantemente, a função da orelha. A parte mecânica da audição (a orelha) é representada por 3 secções: Orelhas externa, média e interna ,em detalhes nas Figuras 4, 5.

A orelha externa é a combinação das cartilagens da aurícula e do meato acústico externo. Já a orelha média inclui a membrana timpânica e outras partes responsáveis por desempenhar seu funcionamento, que são principalmente os três ossos, chamados de

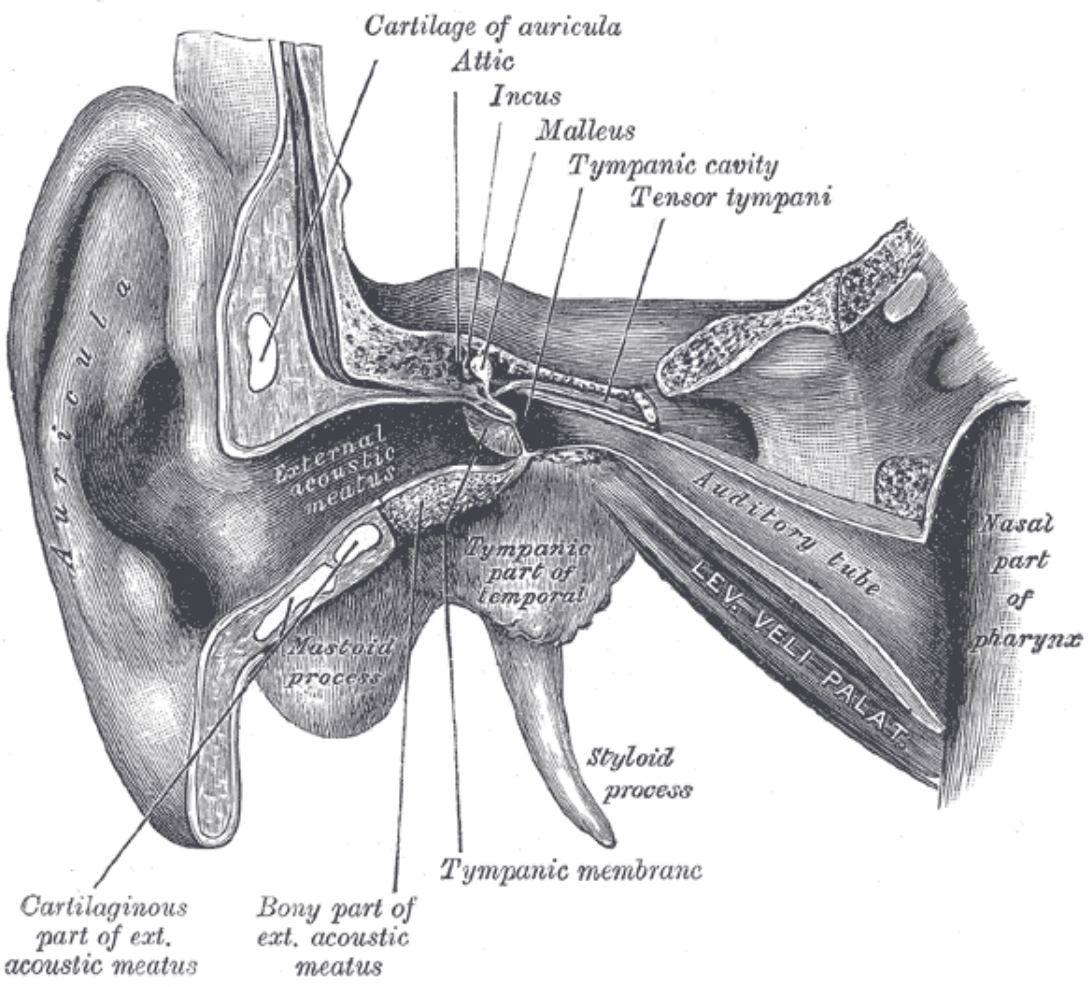


Figura 4 – Orelha Média e Orelha Externa.

Fonte: [Gray & Lewis 1918]

martelo, bigorna e estribo, que transferem o movimento da membrana induzido pelas ondas sonoras que passam e são ampliadas pelo ouvido externo.

A vibração é transmitida da membrana timpânica para os ossos e para a orelha interna através da janela oval da cóclea.

Finalmente, chegamos à orelha interna, em detalhes na Figura 5. Ela é composta pela cóclea, uma cavidade com formato de caracol e três canais semicirculares chamados de ampolas anterior, superior e posterior. A cóclea é preenchida com um fluido que é excitado pelo movimento do estribo na entrada na extremidade da cóclea chamada coclhlear fenestra ovalis. O movimento do estribo induz ondas de pressão no fluido da orelha interna que excitam milhares de cílios dentro do espiral da cóclea (scala tympani). Os cílios são agrupados em quatro linhas, sendo que uma delas está dentro do espiral, perto do centro da curvatura. Essa linha é conectada ao nervo auditivo e transmite o sinal para o cérebro para a cognição. As outras três linhas no outro extremo recebem um feedback do cérebro, que permite a pré-amplificação do movimento do fluido. O formato

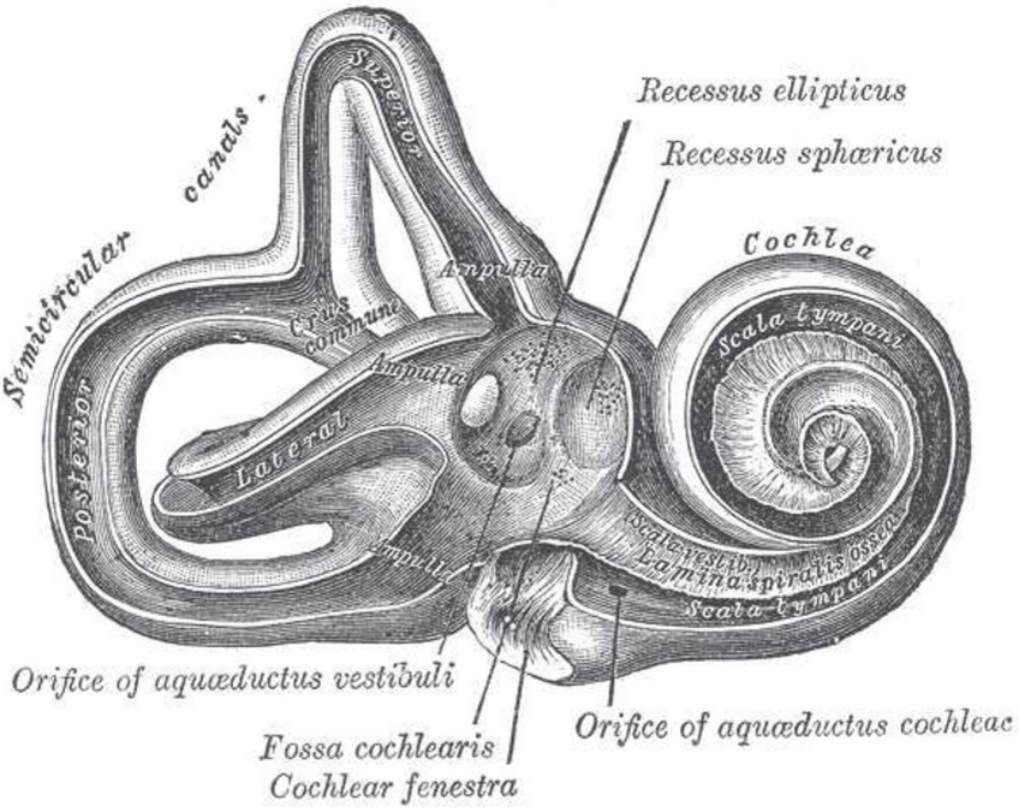


Figura 5 – Orelha interna.

Fonte: [Gray & Lewis 1918]

espiral da scala tympani proporciona uma habilidade cognitiva logarítmica do som, que é muito importante para o desenvolvimento de modelos de falantes e extração atributos do sinal de voz. Começando no coclhlear fenestra ovalis, o áudio de alta frequência é processado. À medida que o som viaja em direção ao helicotrema, os componentes de alta frequência são suprimidos e somente os componentes de tons mais baixo sobrevivem de tal forma que, próximo ao helicotrema, apenas o tom mais baixo é percebido.

Uma vez que os cílios estão excitados, o sinal que eles geram é transportado através do feixe do nervo auditivo para o córtex auditivo localizado nos hemisférios direito e esquerdo do cérebro.

2.2 ALGORITMOS DE CLASSIFICAÇÃO

Muitas aplicações de relevância prática em inteligência artificial são baseadas na criação de modelos computacionais contendo conhecimento empregado por um especialista humano. A tarefa de classificação é feita desta forma, sendo que os objetos a serem classificados são descritos por um conjunto de atributos. Cada objeto deve ser associado a uma classe, dentre o conjunto de classes possíveis para o problema. Os atributos são variáveis observáveis contínuas ou categóricas. A classe é uma variável dependente, cujo

valor é associado aos atributos [Von Zuben 2010]. A classe será sempre uma variável dependente categórica para problemas de classificação.

A construção de modelos computacionais de classificação é feita geralmente a partir de informações de especialistas no problema ou a partir de observações de treinamento de objetos rotulados visando a identificação de relacionamentos entre as variáveis dependentes e independentes na base de observações.

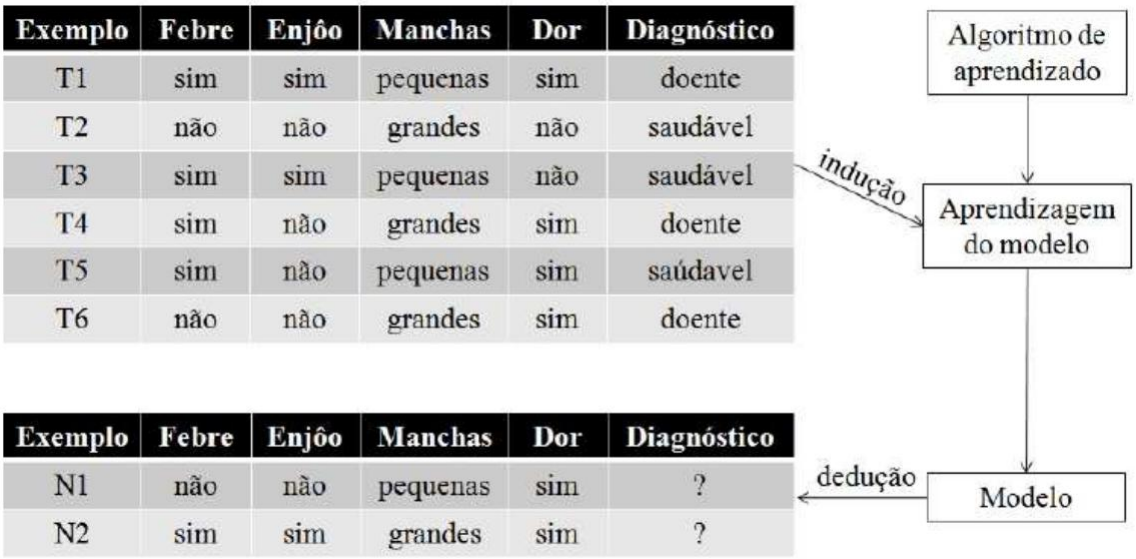


Figura 6 – Indução de um classificador e dedução das classes para novas amostras

Fonte: [Von Zuben 2010]

Um exemplo de problema de classificação é o diagnóstico médico, ilustrado na Figura 6, no qual cada paciente é definido por atributos como febre, enjoo e machas. O objetivo do classificador é fazer o mapeamento dos atributos de forma a induzir um aprendizado do modelo, e assim deduzir o diagnóstico (variável dependente), que contém as classes doente ou saudável.

2.2.1 Árvores de Classificação

Árvores de classificação são métodos de classificação baseados em árvores. Esses métodos se baseiam em segmentar, ou estratificar, o espaço do preditor em um número de regiões simples. O espaço preditor é um espaço multi-dimensional que contempla todos os valores possíveis dos atributos que descrevem a observações dadas. Deste modo, para realizar uma predição com uma dada observação, tipicamente é utilizada a média ou moda das observações de treinamento nas regiões às quais elas pertencem criando regras. Por exemplo em um problema de classificação de uma dada doença, caso o espaço do preditor for segmentado a partir do atributo idade, formando duas regiões, uma com idade maior que certo valor e outra com idade menor, a predição de uma nova observação que se

encaixar em uma das regiões, será classificada utilizando a moda ou a classe majoritária das observações de treinamento que contemplavam a mesma região do espaço do preditor. Como o conjunto de regras para a segmentação do espaço do preditor pode ser resumido em uma estrutura de árvore, podemos considerar esse tipo de abordagem como uma árvore de decisão [James *et al.* 2013].

A tarefa de construir uma árvore de decisão com objetivo de classificar uma observação é data pela segmentação binária do espaço de predição com base na maior ocorrência desta classe em relação às amostras de treinamento na mesma região [James *et al.* 2013]. Ou seja, é utilizado um algoritmo recursivo de busca gulosa que, procura sobre um conjunto de atributos, aqueles que dividem melhor o conjunto de observações em subconjuntos, começando inicialmente com um único nó, chamado raiz [Von Zuben 2010]. A Figura 7 representa um exemplo de árvore de decisão em problemas de classificação, onde as folhas indicam as classes (cliente com risco e cliente sem risco) e os nós de decisão que definem regras ou testes sobre algum valor de um atributo ou mais atributos formando uma sub-árvore para cada um dos valores possíveis desta regra.

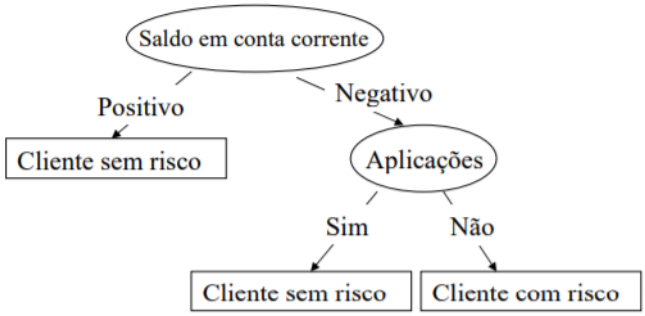


Figura 7 – Exemplo fictício de árvore de decisão, tomando atributos de clientes de uma instituição financeira.

Fonte: [Von Zuben 2010]

Os critérios para selecionar atributos em cada nó da árvore que irão particioná-la é uma tarefa improntante na construção de uma árvore de classificação. Os critérios utilizados são definidos em termos da distribuição da classe de treinamento antes e depois da divisão [Von Zuben 2010].

O particionamento de uma árvore é realizado com objetivo de obter maior grau de pureza, ou seja, um particionamento que resulta em nós com predominantemente mesma classe. A medida deste grau de pureza pode ser feita de diversas formas como, por exemplo, utilizando o índice Gini, o qual é utilizado no algoritmo CART (do inglês *Classification and Regression Trees*) de construção de árvores de decisão [Breiman *et al.* 1984].

O índice Gini é representado pela Expressão 2.1:

$$Gini = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{2.1}$$

onde $\hat{p}_{mk}$ representa a proporção das observações de treinamento na região m que pertence à classe k.

O índice Gini representa uma media da variância total entre as K classes. O índice tende a 0 se $\hat{p}_{mk}$ se aproxima de 0 ou 1 [James *et al.* 2013]. A partir deste valor, os atributos são escolhidos para particionar o espaço de predição e enfim criar a árvore de de classificação.

2.2.2 Métodos de agregação de árvores de decisão

Árvores de classificação permitem boa interpretabilidade na medida que utiliza da criação de regras em uma estrutura de árvore de decisão. Por outro lado, não tem o mesmo desempenho em acurácia que outros métodos de classificação. Nesta seção, serão apresentados brevemente dois exemplos de métodos de agregação de árvores de decisão: RF (*Random Forest*) e GB (*Gradient Boosting*).

Árvores de decisão são considerados classificadores com alta variância. Isso significa que se dividirmos as observações de treinamento em amostras aleatórias e criarmos uma árvore de decisão para cada uma, o resultado pode ser bem diferente. Já classificadores com baixa variância, como a regressão logística produzem resultados semelhantes, mesmo com dados diferentes em cada um [James *et al.* 2013].

O algoritmo RF utiliza um método de reamostragem chamado *bootstrap* com objetivo de reduzir o problema de alta variância das árvores de decisão. O RF utiliza o conjunto de treinamento e cria subconjuntos com observações aleatórias e com reposição. Nestes conjuntos, também são selecionados números diferentes de atributos e assim várias árvores de decisão são construídas em paralelo. A predição de cada uma das árvores aleatórias é contabilizada e a classe mais frequente é selecionada como predição para o classificador RF [James *et al.* 2013].

O classificador GB utiliza árvores de decisão criadas de forma a obterem um melhor desempenho assim como o RF, mas com algumas diferenças. O GB utiliza uma arquitetura de árvores de decisão em paralelo, onde cada árvore de decisão treina utilizando os erros da árvore anterior. O peso de cada uma das árvores no processo de predição é dado por um processo de otimização na etapa de treinamento, utilizando o algoritmo gradiente descendente [Friedman 2001].

2.3 EXTRAÇÃO DE CARACTERÍSTICAS DA VOZ

Com o objetivo de desempenhar a tarefa de classificação a partir de dados contidos em um sinal de voz, é necessário extrair características deste sinal. Para obter informações relevantes do sinal de voz, serão explicados nesta seção técnicas para obtenção de características importantes da voz, que possam ser aprendidas pelo classificador, passando desde a pré-ênfase do sinal, até a extração dos MFCCs.

2.3.1 Pré-Ênfase do sinal

O sistema auditivo humano possuí a cóclea na orelha interna, a qual utiliza de um mecanismo, junto com o cérebro, para ampliar a percepção à diferentes bandas de frequência. Uma vez que desejamos detectar a voz de forma automática, podemos fazer algo similar para extrair mais informações de altas frequências, como sons fricativos, e informações perdidas devido ao mecanismo de produção da fala por humanos, a partir de um filtro de pré-ênfase [Beigi 2011]. Sua função de transferência é descrita na expressão 2.2:

$$H_p(z) = 1 - \alpha z^{-1}, \tag{2.2}$$

onde α é o parâmetro de pré-ênfase. Utiliza-se na maioria das aplicações de processamento de voz $0.95 \leq \alpha \leq 0.97$ [Beigi 2011].

Esse filtro realiza uma equalização de bandas de frequência, fazendo com que aumente a energia do sinal em altas frequências, fazendo com que o modelo tenha disponível essa informação, como ocorre em sons de consoantes fricativas, como [S] e [Z]. As Figuras 8 e 9 mostram a diferença entre um sinal de voz com e sem a pré-ênfase. Note que a distribuição entre as amplitudes do sinal foi alterada na Figura 8, enquanto na imagem 9 é possível notar a queda na potência do sinal em altas frequências.

2.3.2 Quadros e Janelas

A separação do sinal de voz em quadros é essencial para a extração de características de um sinal de voz. Uma das técnicas mais utilizadas é a divisão do sinal não estacionário em uma sequência de pequenos quadros de 20 a 40 ms [Elton, Vasuki & Mohanalin 2016], como pode ser visto na Figura 10. Quadros pequenos são importantes pois o sinal de voz é não estacionário, ou seja, é um sinal cujos parâmetros estatísticos variam em função do tempo. Essa característica faz com que seja razoável segmentar porções menores do sinal e obter características de sinal estacionário neste local.

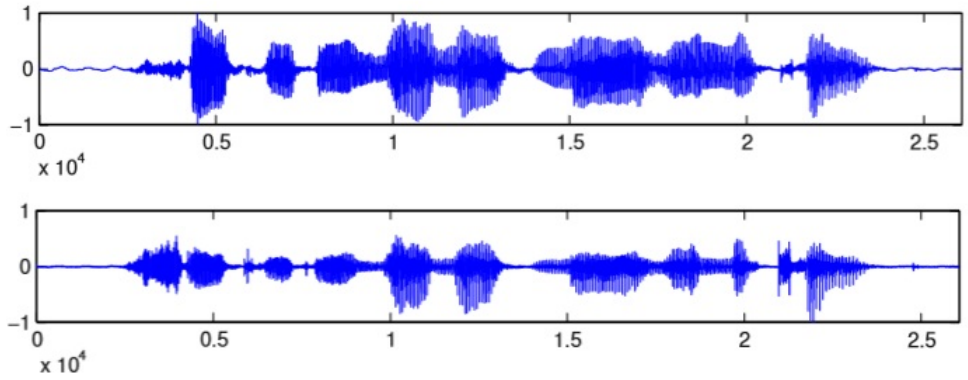


Figura 8 – Sinal de voz sem pré-ênfase (acima) e com pré-ênfase (abaixo).

Fonte: [Beigi 2011]

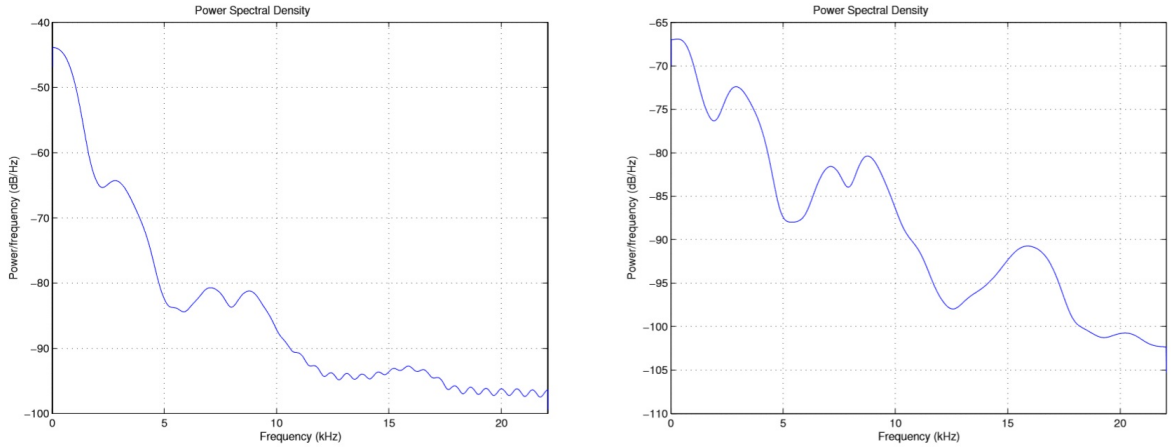


Figura 9 – Análise da densidade espectral de potência do sinal de voz sem pré-ênfase (esquerda) e com pré-ênfase (direita).

Fonte: [Beigi 2011]

O tamanho do quadro é importante para a troca entre resolução no domínio do tempo e da frequência. Se o quadro é muito grande, ele não será capaz de capturar propriedades locais no domínio do tempo, entretanto, se for muito pequeno, a resolução no domínio de frequência será perdida [Nijhawan & Soni 2013], sendo possível perder características de fonemas de duração maior que a duração dos quadros, uma vez que eles não caberiam no mesmo quadro [Beigi 2011].

Os quadros são sobrepostos em 25% a 75% do seu próprio tamanho, como mostra a Figura 10 com 50% de sobreposição. O motivo para isso é a aplicação das janelas de Hamming, que faz com que haja perda de informação no começo e no final de cada quadro. Assim, a sobreposição irá recuperar parte da informação perdida [Nijhawan & Soni 2013].

O processo de janelamento é a multiplicação de N amostras do sinal por uma função janela. O janelamento produzido por janelas de Hamming é a técnica mais utilizada

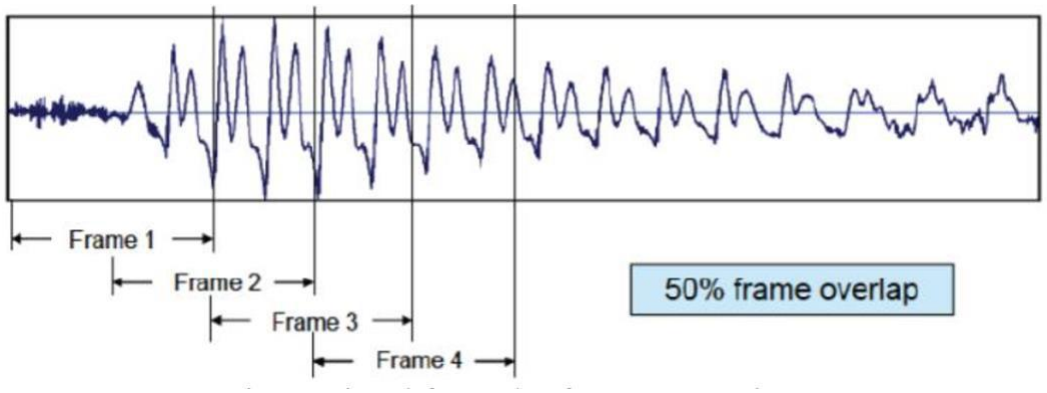
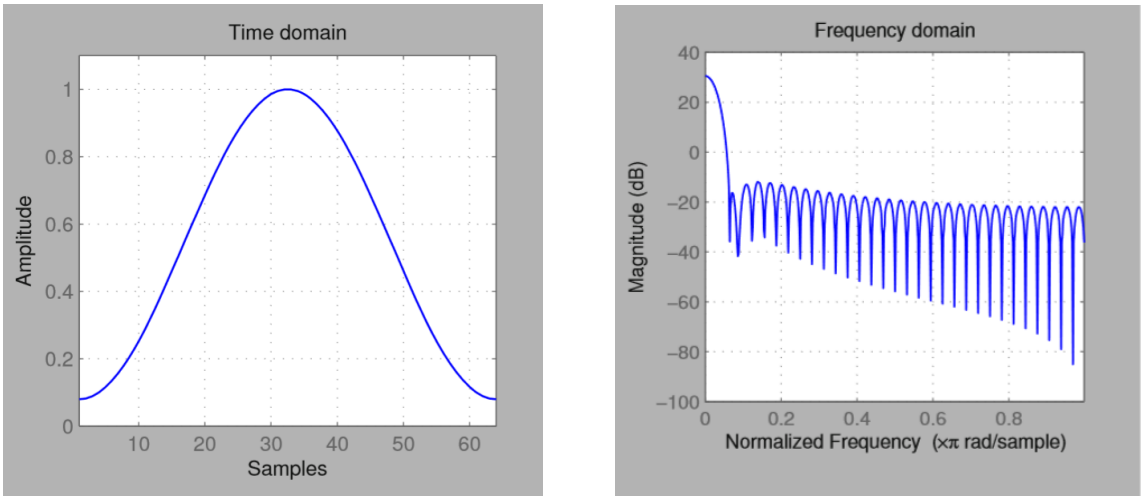


Figura 10 – Processamento de voz dividindo em quadros (frames).

Fonte: [Nijhawan & Soni 2013]

em sistemas de reconhecimento de fala [Zhonghua & Rongchun 2003]. Sua definição é representada pela expressão 2.3, e pela Figura 11

$$W_H(n) = 0.54 - 0.46 \times \cos\left(\frac{2n\pi}{N-1}\right) \tag{2.3}$$



(a) Domínio do tempo (b) Domínio da freqência

Figura 11 – Janela de Hamming com 64 amostras.

Fonte: [Beigi 2011]

As janelas são aplicadas para evitar descontinuidades antinaturais na voz, nos quadros, e distorções no espectro subjacente [Tiwari 2010, Hasan *et al.* 2004]. Uma boa função de janela possui um lóbulo principal estreito e baixos níveis de lóbulo lateral em sua função de transferência [Nijhawan & Soni 2013]. A janela de Hamming possui essa característica, como pode ser visto no item b da Figura 11.

2.3.3 MFCC

Mel frequency cepstral coefficients (MFCC), é provavelmente o mais conhecido e utilizado método para extrair informação em técnicas de reconhecimento de fala e do falante. Mel é uma unidade de medida baseada na percepção de frequência do ouvido humano. A escala mel tem espaçamento aproximadamente linear antes dos 1000 Hz e logarítmica acima deste valor [Nijhawan & Soni 2013]. A aproximação da forma mel de frequência pode ser expressa da seguinte forma:

$$mel(f) = 2596 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.4}$$

onde a variável f é a frequência real e a $mel(f)$ a frequência percebida.

O método de obtenção dos MFCCs pode ser exemplificado em 5 passos, conforme o diagrama de blocos abaixo, representado pela Figura 12:

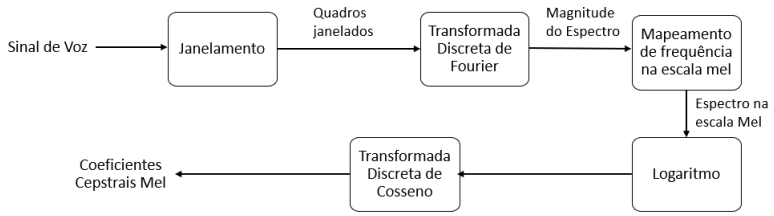


Figura 12 – Diagrama de blocos para obtenção dos MFCCs para um sinal de voz.

Fonte: [Tiwari 2010]

De acordo com o diagrama de blocos da Figura 12, a obtenção dos MFCCs pode ser descritas pelas seguintes passos:

1. Transformar o sinal de voz em quadros.
2. Obtenção da transformada de Fourier em tempo discreto de uma janela do sinal;
3. Mapear a potência do espectro obtida acima na escala mel, utilizando janelas triangulares com sobreposição;
4. Calcular o logaritmo da potência de cada frequência na escala mel;
5. Obtenção da transformada discreta de cosseno da lista de potências na escala mel logarítmica, como se fosse o sinal inteiro;

O passo 1 é realizado ao dividir o sinal de voz em quadros sobrepostos, que são multiplicados por uma função janela que por sua vez é processado no passo 2 pela Transformada Discreta de Fourrier (TDF). Esse processo é feito com objetivo de obter o

espectro de cada quadro, fazendo com que seja obtida característica do sinal do domínio da frequência. Esse processo pode ser feito de maneira mais rápida utilizando o algoritmo FFT, do inglês *Fast Fourier Transform*.

O espectro resultante é então mapeado seguindo a escala mel conforme o passo 3. Esse processo ocorre de modo a simular a percepção humana, onde a informação presente em componentes de baixa frequência do sinal são mais importantes. Deste modo, utiliza-se de bancos de filtros triangulares espaçados seguindo a escala mel, como está é ilustrado na Figura 13. O passo 4 é realizado transformando o espectro resultante do banco de filtro para a escala logarítmica.

O passo 5 é feito utilizando a transformada discreta de cosseno (DCT) do logaritmo do espectro de cada um dos filtros do banco de filtros triangulares, resultando no coeficiente mel cepstral, como mostra a expressão 2.5 [Huang *et al.* 2001]

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left(\pi n(m + 1/2)/M\right), \qquad 0 \leq n < M \tag{2.5}$$

onde M representa o banco de filtros ($m = 1, 2, ..., M$), m é o índice de cada filtro, $S[m]$ é o logaritmo da energia de cada filtro e n é o número de coeficientes desejado. Usualmente, utiliza-se n = 13 [Huang *et al.* 2001].

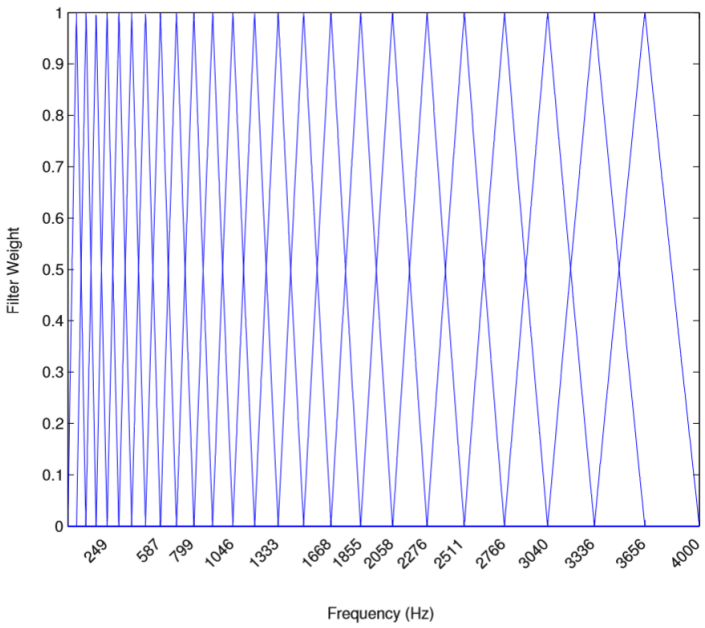


Figura 13 – Banco de filtros triangulares na escala mel para um sinal com taxa de amostragem 8 kHz.

Fonte: [Beigi 2011]

2.4 VAD

A Detecção de Atividade de Voz (VAD) é a técnica de processamento de fala que discrimina regiões de fala e não fala. Silêncio, ruído ou outras informações acústicas não relacionadas são tratados como regiões sem fala. O desafio da VAD é detectar fala em ambientes com uma relação sinal-ruído (SNR) baixa e também sob a influência de ruídos não-estacionários que causam erros significativos [Elton, Vasuki & Mohanalin 2016].

A Figura 14 exemplifica a aplicação de VAD utilizando uma árvore de decisão para obter a classificação de quadros contendo voz. Os atributos utilizados são representados por características da voz como energia e o *Pitch* (quantidade percebida do sinal que está relacionada com a frequência fundamental de vibração dos cordas vocais durante determinado tempo [Beigi 2011])

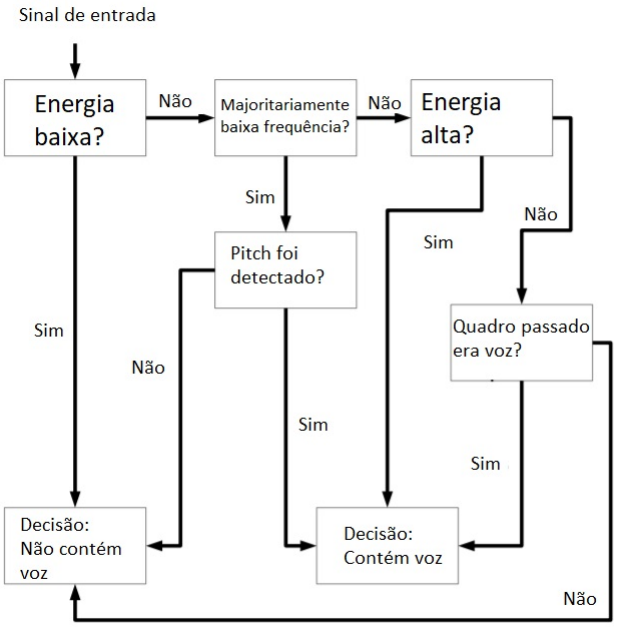


Figura 14 – Árvore de decisão aplicada ao problema de VAD.

Fonte: [BÄckstrÖm 2016]

A presente seção tem como objetivo expor o tema principal do projeto, VAD em ambientes ruídosos e expor algumas aplicações que o utilizam para melhorar sua performance.

2.4.1 VAD em ambientes ruidosos

Um grande problema em diversas áreas do processamento de voz é a determinação da presença de períodos de fala, dado a presença de ruído no sinal. A classificação de um período de voz deixa de ser trivial conforme o nível de ruído de externo aumenta. A

Figura 15 mostra o desafio de detectar voz conforme o ruído aumenta e mascara o sinal todo. A seleção de atributos corretos para uma classificação de voz robusta ao ruído é uma tarefa desafiadora [Ramirez, Górriz & Segura 2007].

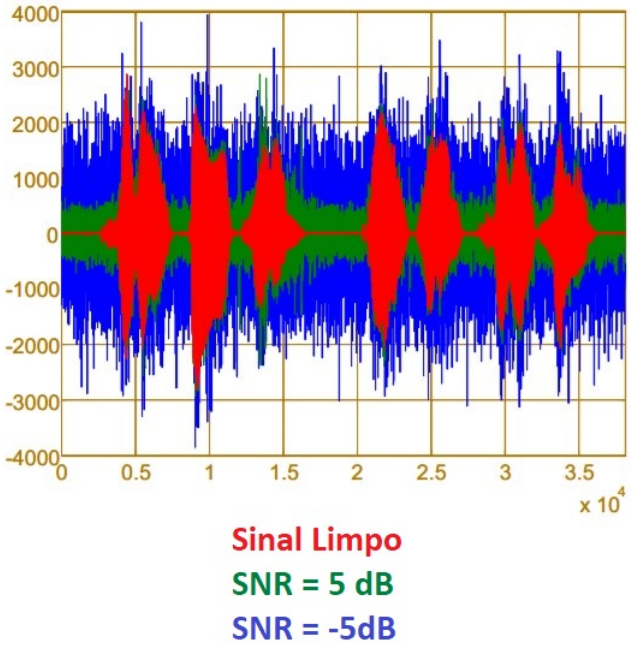


Figura 15 – Sinal de voz com diferentes SNRs.

Fonte: [Ramirez, Górriz & Segura 2007]

VAD requer o estágio de extração de características, usualmente utilizando características acústicas. Métodos tradicionais utilizam um limiar de energia, taxa de cruzamento zero (ZCR), diferença de energia entre quadros contendo fala dos que não contém, características utilizando coeficientes wavelet, distância ceptral e periodicidade da fala. Todos com sua eficiência limitada pelo ruído, representado pela SNR (relação sinal-ruído) [Elton, Vasuki & Mohanalin 2016].

2.4.2 Aplicações VAD

A técnica VAD na presença de ruídos fortes e dinâmicos é um problema que tem chamado a atenção de pesquisadores recentes da área. Sua utilização é vital para modelos modernos de processamento de sinais de voz. A seguir, são apresentadas brevemente algumas áreas de aplicação de VAD.

2.4.2.1 Reconhecimento de Falante

A tarefa de reconhecimento de falante se baseia no processo de reconhecer automaticamente o falante, baseado nas características contidas no sinal de voz. A maioria dos sistemas de reconhecimento de falante são feitos em duas fases importantes, extração

de atributos da voz e mapeamento do falante [Nijhawan & Soni 2013]. A primeira se baseia na extração de características presentes no sinal de voz que possam caracterizar indivíduo e posteriormente identifica-lo através delas. A segunda etapa consiste em utilizar características extraídas de um falante e compara-las às características guardadas de outros falantes com objetivo de reconhecer o falante.

Estudos na técnica de reconhecimento de falante estão aumentando, resultando na maior utilização de algoritmos de aprendizado de máquina, que também podem ser utilizados para o VAD. A Figura 16 exemplifica um trabalho onde o VAD foi utilizado para reconhecer o falante. Ele é utilizado após a extração dos atributos e auxilia como um filtro de segmentos do sinal de voz capturado por um celular que contenha fala, podendo melhorar o desempenho tanto em relação à precisão como em economia de processamento [Wahib-Ul-Haq 2015].

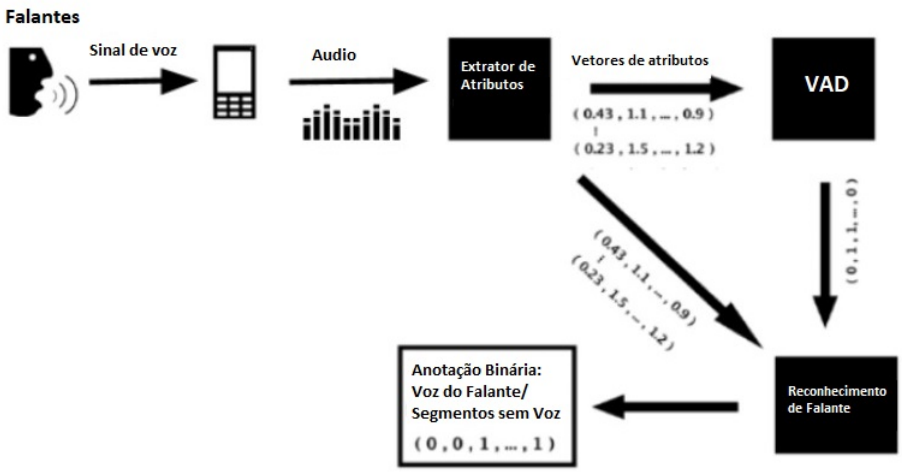


Figura 16 – Aplicação VAD em modelo de reconhecimento de falante utilizando telefone celular.

Fonte: [Wahib-Ul-Haq 2015]

2.4.2.2 Aprimoramento da Voz

O aprimoramento da voz foca em melhorar o desempenho de sistemas de comunicação em ambientes ruidosos. Na maioria dos casos, foca em suprimir o ruído de fundo de um sinal com ruído. Algumas técnicas, como o filtro de Wiener [Chirtmay, Tahernezhadi *et al.* 1997] requerem a estimativa do espectro de potência do sinal com fala sem ruído e do espectro de potência do ruído. A estimativa pode ser dada com auxílio do VAD [Ramirez, Górriz & Segura 2007].

2.4.2.3 Reconhecimento de Fala

Reconhecimento de fala é o processo de converter um sinal de voz em uma sequência de palavras por meio de um algoritmo implementado como um programa de computador.

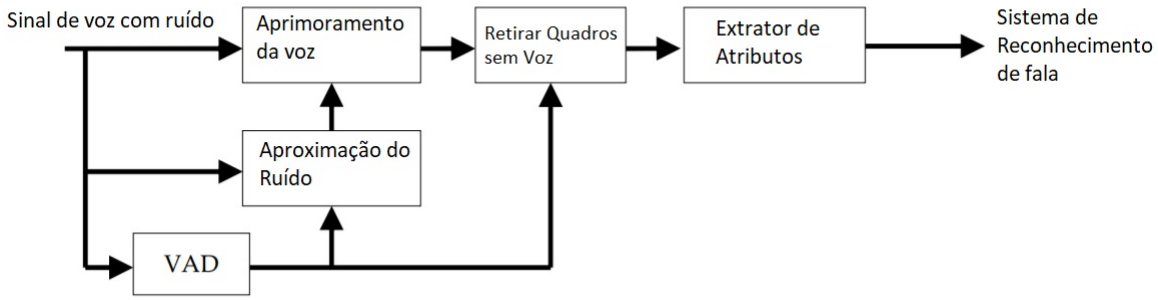


Figura 17 – Diagrama de blocos para um sistema de reconhecimento de fala com utilização do VAD.

Fonte: [Ramirez, Górriz & Segura 2007]

Essa tecnologia tornou possível para um computador obedecer comandos humanos por voz e entender idiomas [Anusuya & Katti 2010].

O desempenho de um modelo de reconhecimento de fala é fortemente influenciado pela qualidade do sinal de voz. A técnica de VAD é muito útil para melhorar a performance desses modelos. O VAD é frequentemente utilizado em sistemas de reconhecimento de fala, retirando os segmentos de áudio que não contém fala, e auxiliando em modelos de aprimoramento de voz, como mostra a Figura 17, de modo a reduzir o erro no reconhecimento de palavras faladas em um áudio [Ramirez, Górriz & Segura 2007].

3 METODOLOGIA

3.1 BASE DE DADOS

A base de dados utilizada é livre e foi utilizada no projeto de [Kim & Hahn 2018], disponível em [Kim 2017]. Os nomes dos áudios presentes nesta base estão contidos na tabela 1, junto com suas SNRs.

Tabela 1 – Tabela contendo os áudios e suas SNRs

Áudios	Bus Stop	Park	Room
Duração (min)	30.02	30.07	30.05
SNR Média (dB)	5.61	5.71	18.26
Tempo Falado (%)[1]	40.12	26.85	30.44

O treinamento foi realizado com o áudio Bus Stop, enquanto os áudios Park e Room foram utilizados para teste e para comparar o desempenho de diferentes metodologias de VAD em ambientes diferentes. O áudio Bus Stop foi gravado em um ponto de ônibus, com bastante ruído de trânsito, assim como a base Park, gravada em um parque. Room, por sua vez tem menor ruído, gravada em um escritório.

O dispositivo utilizado para gravação foi o smartphone Samsung Galaxy S8. Áudio em diferentes ambientes foram gravados, com dois falantes coreanos do sexo masculino. As marcações de atividade de voz foram anotadas manualmente. Pelo fato desta base de dados ter sido criada em um ambiente real, ruídos de diferentes fontes são incluídos, como choros de bebê, barulho de insetos e cliques de mouse.

Os arquivos de áudio foram transformados utilizando o software SoX de uma taxa de amostragem de 16 kHz para 8 kHz e de 32 para 16 bits de profundidade por amostra utilizando o formato PCM (Modulação por código de pulsos). Essa transformação tem objetivo de diminuir a qualidade do áudio, e torná-lo semelhante aos padrões utilizados em bases de dados de áudios em sistemas de telefonia [Daengsi *et al.* 2012]. Já que que o objetivo deste projeto é detectar voz humana em ambientes ruidosos e com áudios de baixa qualidade.

3.2 ARQUITETURA DO SISTEMA

O sistema apresentado se propõe a realizar a tarefa de detectar segmentos do sinal de voz contendo fala. Deste modo, foi proposta uma arquitetura que funciona em cinco passos, descritos a seguir e ilustrados pela Figura 18:

[1] Porcentagem de tempo em que há voz no áudio

1. Redução do ruído aplicando filtro de pré-ênfase;
2. Separar o sinal em quadros;
3. Extrair características do sinal na forma de MFCCs.
4. Utilizar as características de cada quadro para treinamento ou teste do classificador escolhido;
5. A resposta do classificador para quadro é utilizada para obtenção de trechos de fala e não fala contidos no sinal.

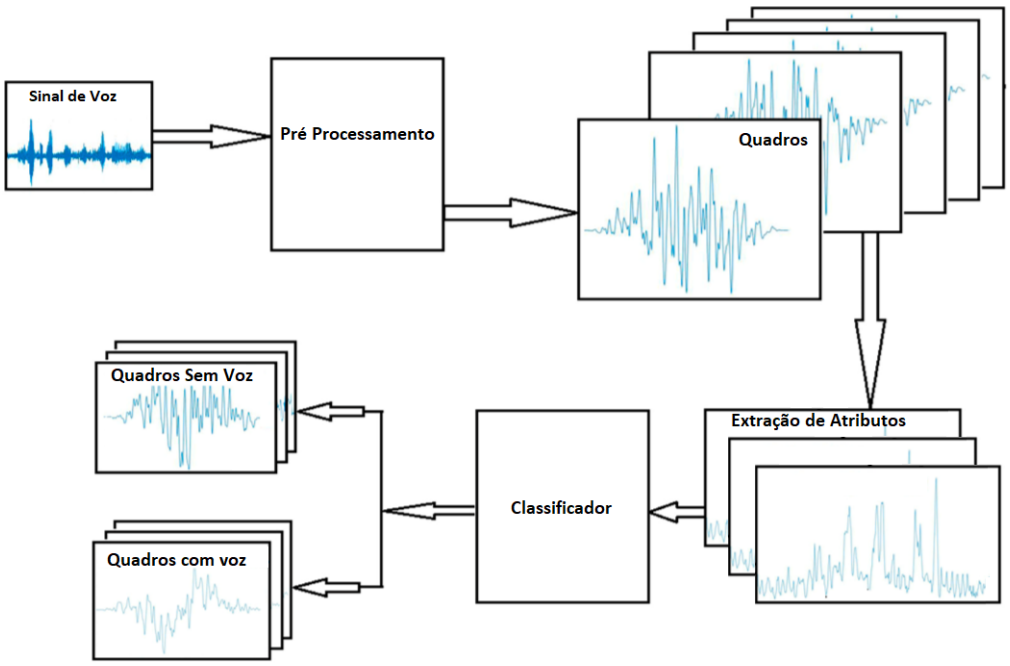


Figura 18 – Arquitetura proposta para realização do projeto.

Fonte: Adaptado de [Elton, Vasuki & Mohanalin 2016]

O passo 1 foi feitoa partir da pré-ênfase do sinal utilizando um filtro passa alta com parâmetro α de 0.97, conforme explicado no na secção 2.4.1. O passo 2 foi implementado com objetivo de separar o sinal em quadros de tamanho 25 ms e sobreposição de 15 ms por frame. Como o sinal tem uma frequência de amostragem de 8 kHz, temos cada frame com 200 amostras e o número de frames por sinal dependera do tamanho do mesmo. A extração das características do sinal foi feita no passo 3, utilizando apenas os coeficientes obtidos a partir da extração dos MFCCs. Os passos 1,2 e 2 foram realizados com a utilização biblioteca Python Speech Features [Python Speech Features 2018] sem alterar seus demais parâmetros.

O classificador descrito no passo 4 utiliza como entrada os atributos extraídos de cada quadro do sinal para treinamento e no passo 5 faz a inferência, com objetivo de distinguir quadros com ou sem voz, conforme apresentado na figura 18. O projeto será realizado com a utilização de um classificador. Para este projeto serão utilizados 2 classificadores diferentes, Random Forest (RF) e Gradient Boosting (GB), ambos da biblioteca Scikit-Learn [Pedregosa *et al.* 2011] em Python.

Após a classificação, os quadros classificados de forma binária entre fala e não fala passaram por um processo de suavização, utilizando um filtro de mediana. O filtro me mediana baseia-se em replicar a resposta dos do quadro vigente com a mediana da resposta do classificador para os quadros vizinhos. A vizinhança utilizada neste projeto foi de 51 quadros, ou seja a resposta de cada quadro é a mediana da resposta dos 25 quadros anteriores e 25 quadros posteriores ao quadro vigente. VADs que utilizam decisão quadro a quadro, podem utilizar suavização para aumentar a robustez contra ruído. A motivação para essas técnicas é encontrada no processo de produção da fala. Esse problema ocorre pelo fato da pronuncia de algumas palavras terem trechos com baixa energia ou até mesmo algum intervalo pequeno sem voz. Deste modo, a suavização pode estender também períodos de fala mascarados por ruídos acústicos [Ramirez, Górriz & Segura 2007], como mostra a Figura 19.

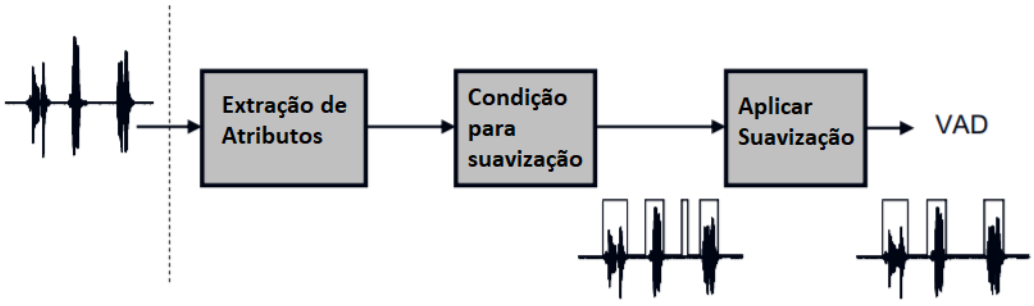


Figura 19 – Exemplo esquemático da técnica de suavização para VAD.

Fonte: Adaptado de Ramirez, Górriz & Segura 2007

4 RESULTADOS E DISCUSSÃO

4.1 MÉTRICAS DE DESEMPENHO EM ALGORITMOS DE CLASSIFICAÇÃO

Há diversas formas de medir a performance de modelos de classificação. Uma das formas é partindo de uma matriz de confusão, que tem a informação das observações classificadas corretamente e incorretamente por classe. A tabela 2 apresenta um exemplo de matriz de confusão para duas classes (Positivo ou Negativo), onde VP representa verdadeiro positivo, FP como falso positivo, FN como falso negativo e VN como verdadeiro negativo.

Tabela 2 – Matriz de Confusão

Classe \Predito	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

A acurácia representada pela Expressão 4.1, é uma das métricas que não leva em conta a distinção das classes, sendo uma métrica não indicada para problemas onde há desbalanceamento entre classes ou quando existe a necessidade de analisar melhor a classificação de cada classe separadamente [Sokolova, Japkowicz & Szpakowicz 2006].

A precisão e o recall, expressões 4.2 e 4.3, respectivamente, levam em conta apenas uma das classes. Tendo como referência os valores positivos, a precisão representará a proporção das observações corretamente classificadas como positivo. O recall ou TVP (Taxa de Verdadeiros Positivos) representa a proporção das observações classificadas como positivo, em relação ao total de observações positivas. A TFP (Taxa de Falsos Positivos), representa a razão entre as observações classificadas erroneamente como positivas em relação à todas as amostras negativas [Sokolova, Japkowicz & Szpakowicz 2006].

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.1}$$

$$Precisão = \frac{VP}{VP + FP} \tag{4.2}$$

$$Recall = \frac{VP}{VP + FN} = TVP \tag{4.3}$$

$$TFP = \frac{FP}{FP + VN} \tag{4.4}$$

A curva ROC (do inglês *Receiver Operating Characteristics Curve*) representa outra forma de analisar um classificador, representada pelos valores de TVP no eixo vertical e TFP no eixo horizontal em um gráfico de duas dimensões. Levando em conta o valor das probabilidades do classificador binário para determinada classe. A área sob a curva ROC, chamada de AUC (do inglês *Area Under the ROC Curve*) consegue medir a performance do classificador em problemas com dados de classes desbalanceadas e medir o custo representado pelo TFP e o benefício TVP baseado em alguma aplicação. A Figura 20 exemplifica a curva ROC [Huang & Ling 2005, Sokolova, Japkowicz & Szpakowicz 2006].

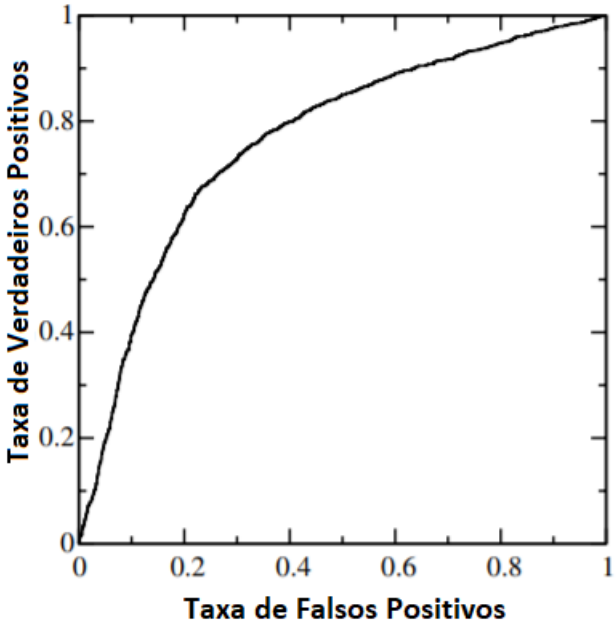


Figura 20 – Exemplo Curva ROC.

Fonte: [Fawcett 2006]

As métricas utilizadas para medir a performance dos modelos foram precisão, recall, acurácia e AUC (área sob a curva ROC, do inglês Receiver Operating Characteristic). Vale ressaltar que a importância de cada uma das métricas pode variar dependendo da aplicação. O Recall por exemplo é importante para aplicações onde não se deseja perder nenhum trecho contendo voz, já a precisão faz mais sentido em aplicações onde os trechos classificados como voz não possam ter segmentos sem voz.

4.2 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Os modelos foram treinados utilizando áudio Bus Stop, e para teste foram utilizadas as bases Park e Room, conforme explicado na sessão anterior. A metodologia proposta, utilizando dois classificadores diferentes, com e sem suavização foi comparada com outros dois modelos detectores de voz, webrtcvad [Google 2017] e van2013robust[Segbroeck, Tsiartas & Narayanan

O primeiro feito por desenvolvedores da Google no projeto WebRTC. Já o segundo é uma metodologia utilizando redes neurais profundas, apresentado no congresso Interspeech do ano de 2013. Os dois últimos modelos não precisaram ser treinados e estão disponíveis para utilização.

A Figura 21 exemplifica de forma visual a atuação do modelo VAD proposto, utilizando o classificador RF. Representado em azul podemos ver o sinal original. O resultado esperado pode ser visto com um traço tracejado, enquanto os resultados obtidos com suavização em laranja. A discriminação entre silêncio e voz é feita considerando os valores 0 (sem voz) e 1 (com voz).

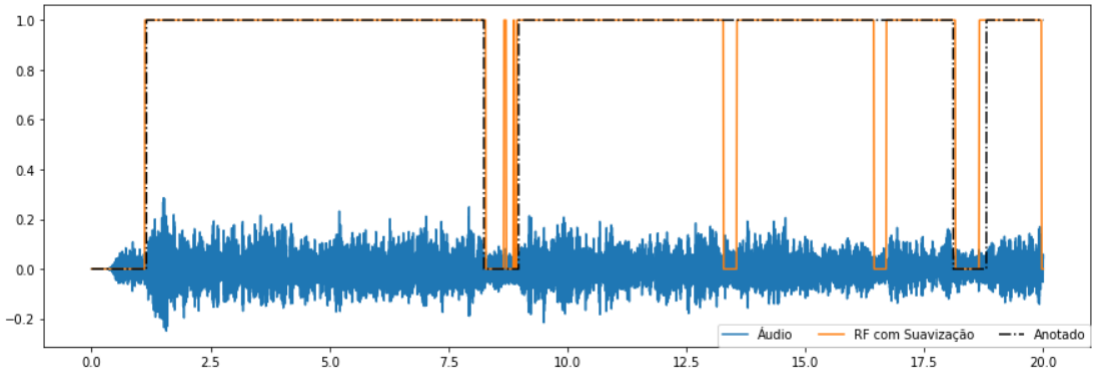


Figura 21 – Exemplo da utilização da técnica proposta VAD, aplicando o classificador Random Forest, com suavização, para um trecho do áudio Park.

O desempenho de cada modelos está exposto na Tabela 3. GB e RF obtiveram melhores resultados com a utilização da suavização para todas as métricas para as bases Park e Room, com exceção do modelo RF para base Room, que apresentou melhora com suavização apenas na métrica Recall.

Os resultados mostraram o modelo RF* como o melhor para o áudio Park, com resultados de 0,95, 0,86, 0,96 para as métricas AUC, acurácia e precisão, respectivamente. O modelo van2013robust, por sua vez foi melhor para o áudio Room, nas métricas, AUC, acurácia e precisão com valores 0,94, 0,92, 0,82, respectivamente, enquanto o melhor recall foi obtido pelo modelo RF*, com valor de 1,00.

Os modelos webrtcvad e van2013robust se mostraram piores que o RF*, para o áudio Park, em todas as métricas, enquanto para o áudio Room, nenhum dos modelos propostos obtiveram resultados melhores. Esse problema pode explicado pelo método de treinamento, utilizando apenas um áudio com uma SNR semelhante ao do áudio Park, mas bem diferente ao do áudio Room. A utilização de maior quantidade de dados em ambientes diferentes poderia melhorar qualidade do treinamento e aumentar a performance do modelo proposto, uma vez que em ambientes com SNR semelhante o modelo proposto

obteve melhor desempenho.

A técnica de suavização por filtro de mediana utiliza nos modelos GB* e RF* mostrou-se importante com melhora na performance em todas as métricas para a base Park. No entanto para a base Room o mesmo não ocorreu apenas para o modelo RF*, que teve apenas a métrica recall melhorada pela utilização desta técnica. Esse resultado pode ser explicado pelo fato da suavização intensificar a má precisão do modelo RF para base Park, transformando os poucos quadros classificados como não voz em quadros de voz.

O modelo GB*, dos modelos treinados para este projeto, foi o melhor no aspecto de generalização. Além de ter seu desempenho melhorado pela utilização da suavização, desempenhou melhor na base Park em relação ao RF*, mostrando que obteve melhor generalização dos dados, enquanto o RF* teve seu desempenho prejudicado quando testado em dados com característica de ruído muito diferentes.

Tabela 3 – Resultados para áudio Park e Room

Modelos	Park				Room			
	AUC	Acurácia	Prec	Recall	AUC	Acurácia	Prec	Recall
GB	0,72	0,67	0,44	0,82	0,72	0,64	0,46	0,92
RF	0,80	0,79	0,58	0,80	0,63	0,50	0,37	0,96
GB*	0,86	0,80	0,58	0,98	0,75	0,66	0,47	0,99
RF*	**0,95**	**0,95**	**0,86**	0,96	0,55	0,38	0,33	**1,00**
webrtcvad	0,77	0,76	0,53	0,78	0,81	0,76	0,57	0,94
van2013robust	0,92	0,89	0,71	**0,99**	**0,94**	**0,92**	**0,82**	0,97

Observações: * representa modelos utilizando suavização por filtro de mediana.

5 CONCLUSÃO

Os resultados obtidos utilizando classificadores baseados em árvores de decisão foram suficientes para desempenhar uma boa extração de trechos com atividade de voz do áudio, mesmo em ambientes bastante ruidosos, principalmente em aplicações em que a métrica Recall seja a mais importante.

Os modelos propostos obtiveram uma redução de desempenho ao serem testados em uma base com SNR muito maior, resultado diferente do que é esperado, uma vez que detecção de voz em ambientes mais ruidosos é uma tarefa mais difícil para os modelos de VAD convencionais baseados na energia do sinal.

Dentre as principais vantagens do método proposto, a criação de um modelo de detecção de voz com resultados consistentes, sem ter necessidade da criação de regras feitas por um ser humano. No entanto, situações onde haja diferentes níveis de ruído podem prejudicar o modelo, que necessita de mais dados de treinamento para obter resultados cada vez mais robustos, em comparação com os modelos utilizados para comparação.

Em linhas gerais, os resultados almejados foram alcançados neste trabalho. Recomenda-se a otimização dos parâmetros dos modelos de classificação, como número de árvores nos métodos de agregação de árvores de decisão, profundidade de cada árvore e taxa de aprendizagem. Além da extração de diferentes atributos além dos MFCCs e obtenção de mais dados rotulados para treinamento para obter melhorias na performance do VAD.

REFERÊNCIAS

Anusuya & Katti 2010 ANUSUYA, MA; KATTI, Shriniwas K. Speech recognition by machine, a review. **arXiv preprint arXiv:1001.2267**, 2010.

BÄckstrÖm 2016 BÄCKSTRÖM, Tom. **Speech Processing: Voice Activity Detection**. Aalto University, 2016. Disponível em: <https://mycourses.aalto.fi/course-/view.php?id=13460>. Acesso em: 17/06/2018.

Beigi 2011 BEIGI, Homayoon. **Fundamentals of speaker recognition**. [S.l.]: Springer Science & Business Media, 2011.

Breiman *et al.* 1984 BREIMAN, L.; FRIEDMAN, J.; STONE, C.J.; OLSHEN, R.A. **Classification and Regression Trees**. Taylor & Francis, 1984. (The Wadsworth and Brooks-Cole statistics-probability series). ISBN 9780412048418. Disponível em: <https://books.google.com.br/books?id=JwQx-WOmSyQC>.

Chirtmay, Tahernezhadi *et al.* 1997 CHIRTMAY, S; TAHERNEZHADI, M *et al.* Speech enhancement using wiener filtering. **Acoustics letters**, v. 21, p. 110–115, 1997.

Daengsi *et al.* 2012 DAENGSI, Therdpong; WUTIWIWATCHAI, Chai; PREECHAYA-SOMBOON, Apiruck; SUKPARUNGSEE, Saowanit. A study of voip quality evaluation: User perception of voice quality from g. 729, g. 711 and g. 722. In: IEEE. **Consumer Communications and Networking Conference (CCNC), 2012 IEEE**. [S.l.], 2012. p. 342–345.

Elton, Vasuki & Mohanalin 2016 ELTON, R Johny; VASUKI, P; MOHANALIN, J. Voice activity detection using fuzzy entropy and support vector machine. **Entropy**, Multidisciplinary Digital Publishing Institute, v. 18, n. 8, p. 298, 2016.

Fawcett 2006 FAWCETT, Tom. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.

Friedman 2001 FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

Gartner 2017 GARTNER. **Gartner's hype cycle for emerging technologies maps the journey**. 2017.

Google 2017 GOOGLE. **Python interface to the WebRTC Voice Activity Detector**. GitHub, 2017. Disponível em: <https://github.com/wiseman/py-webrtcva>. Acesso em: 17/06/2018.

Gray & Lewis 1918 GRAY, Henry; LEWIS, WH. Anatomy of the human body. 20th. **Philadelphia and New York, Lea & Febiger**, 1918.

Hasan *et al.* 2004 HASAN, Md Rashidul; JAMIL, Mustafa; RAHMAN, MGRMS *et al.* Speaker identification using mel frequency cepstral coefficients. **variations**, v. 1, n. 4, 2004.

Huang & Ling 2005 HUANG, Jin; LING, Charles X. Using auc and accuracy in evaluating learning algorithms. **IEEE Transactions on knowledge and Data Engineering**, IEEE, v. 17, n. 3, p. 299–310, 2005.

Huang *et al.* 2001 HUANG, Xuedong; ACERO, Alex; HON, Hsiao-Wuen; REDDY, Raj. **Spoken language processing: A guide to theory, algorithm, and system development**. [S.l.]: Prentice hall PTR Upper Saddle River, 2001.

James *et al.* 2013 JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An introduction to statistical learning**. [S.l.]: Springer, 2013.

Kim 2017 KIM, J. **Voice Activity Detection Toolkit**. GitHub, 2017. Disponível em: <https://github.com/jtkim-kaist/VAD>.

Kim & Hahn 2018 KIM, Juntae; HAHN, Minsoo. Voice activity detection using an adaptive context attention model. **IEEE Signal Processing Letters**, IEEE, 2018.

Nijhawan & Soni 2013 NIJHAWAN, Geeta; SONI, MK. A new design approach for speaker recognition using mfcc and vad. **International Journal of Image, Graphics and Signal Processing (IJIGSP)**, Citeseer, v. 5, n. 9, p. 43–49, 2013.

Pedregosa *et al.* 2011 PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent *et al.* Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011.

Python Speech Features 2018 PYTHON SPEECH FEATURES. **Python Speech Features**. GitHub, 2018. Disponível em: <https://github.com/jameslyons-/python_speech_features>. Acesso em: 20/04/2018.

Ramirez, Górriz & Segura 2007 RAMIREZ, Javier; GÓRRIZ, Juan Manuel; SEGURA, José Carlos. Voice activity detection. fundamentals and speech recognition system robustness. In: **Robust speech recognition and understanding**. [S.l.]: InTech, 2007.

Segbroeck, Tsiartas & Narayanan 2013 SEGBROECK, Maarten Van; TSIARTAS, Andreas; NARAYANAN, Shrikanth. A robust frontend for vad: exploiting contextual, discriminative and spectral cues of human voice. In: **INTERSPEECH**. [S.l.: s.n.], 2013. p. 704–708.

Sokolova, Japkowicz & Szpakowicz 2006 SOKOLOVA, Marina; JAPKOWICZ, Nathalie; SZPAKOWICZ, Stan. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: SPRINGER. **Australasian joint conference on artificial intelligence**. [S.l.], 2006. p. 1015–1021.

Tiwari 2010 TIWARI, Vibha. Mfcc and its applications in speaker recognition. **International journal on emerging technologies**, v. 1, n. 1, p. 19–22, 2010.

Von Zuben 2010 VON ZUBEN, Fernando José. **Árvores de Decisão**. [S.l.]: Departamento de Engenharia de Computação e Automação Industrial, Universidade de Campinas, Campinas, 2010.

Wahib-Ul-Haq 2015 WAHIB-UL-HAQ, Wahib. Speaker detection and conversation analysis on mobile devices. 2015.

Wang, Xu & Li 2011 WANG, Hongzhi; XU, Yuchao; LI, Meijing. Study on the mfcc similarity-based voice activity detection algorithm. In: IEEE. **Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on**. [S.l.], 2011. p. 4391–4394.

Zhonghua & Rongchun 2003 ZHONGHUA, Fu; RONGCHUN, Zhao. An overview of modeling technology of speaker recognition. In: IEEE. **Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on**. [S.l.], 2003. v. 2, p. 887–891.