Pós-Graduação em Ciência da Computação

Hesdras Oliveira Viana

MODELO ADAPTATIVO PARA RECONHECIMENTO DA FALA COM RECONSTRUÇÃO DE CARACTERÍSTICAS AUSENTES

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2017

Hesdras Oliveira Viana

MODELO ADAPTATIVO PARA RECONHECIMENTO DA FALA COM RECONSTRUÇÃO DE CARACTERÍSTICAS AUSENTES

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Orientador: *Aluízio Fausto Ribeiro Araújo*

RECIFE
2017

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

V614m Viana, Hesdras Oliveira
Modelo adaptativo para reconhecimento de fala com reconstrução de características ausentes / Hesdras Oliveira Viana. – 2017.
89 f.: il., fig., tab.

Orientador: Aluízio Fausto Ribeiro Araújo.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2017.
Inclui referências.

1. Inteligência artificial. 2. Reconhecimento de fala. I. Araújo, Aluízio Fausto Ribeiro (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2017-190

Hesdras Oliveira Viana

Modelo Adaptativo para Reconhecimento da Fala com Reconstrução de Características Ausentes

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutora em Ciência da Computação

Aprovado em: 08/05/2017.

_____

Orientador: Prof. Dr. Aluizio Fausto Ribeiro Araújo

BANCA EXAMINADORA

_____

Prof. Dr. Tsang Ing Ren
Centro de Informática / UFPE

_____

Prof. Dr. Hansenclever de França Bassani
Centro de Informática / UFPE

_____

Prof. Dr. Jugurta Rosa Montavao Filho
Departamento de Engenharia Elétrica / UFS

_____

Prof. Dr. Francisco Madeiro Bernardino Junior
Escola Politécnica de Pernambuco / UPE

_____

Prof. Dr. Alexandre Magno Andrade Maciel
Escola Politécnica de Pernambuco / UPE

Dedico esse trabalho aos meus heróis, comumente chamados de pai e mãe, Nivaldo Viana e Miralva Santos, aos meus conselheiros, comumente chamados de irmãos, Rondinelli Viana e Nivaldo Júnior, as minhas tias, em especial Valda Santos, e a minha esposa, Karla Abobreira. A todos eu dedico.

Agradecimentos

Realização Sim, Gratidão Sempre.
O que me trouxe a essa Universidade?
A escolha foi durante a graduação.
Com meu orientador professor doutor Roque Mendes Prado Trindade.
Aqui tornei-me Mestre. E, no momento, chegando a Doutor.
Recebi uma visão diferenciada do saber.
Respeito e confiança a coragem de fazer e vencer.
Sou simplesmente renascente. Relevante, renomado.
Eficácia comprovada da vontade desejada.
Talvez o modo de ser, quem sabe de amigos
Encarando, não sei como, os problemas com prazer.
Gostar é pouco demais no que se quer realizar
É preciso se permitir, buscar o sim e o não
E então se apaixonar.
Às vezes frustrado e limitado. Receoso de fazer ou dizer
De se achar e cometer enganos, parar em meio à indecisão
E dizer a si mesmo que: A graça das graças é nunca desistir.
Percebo que não se nasce, uma única vez.
A cada vitória ou conquista, se nasce.
Nasce na amizade sincera.
Nasce com a esperança do esperado sucesso.
Encontrei na Universidade Pessoas fascinantes.
Entre elas, o professor doutor e orientador Aluízio Fausto Ribeiro Araújo
Pessoa desapegada do egoísmo acadêmico. Sábio, um Gênio!
Transformou o meu desencanto
A minha mente preocupante.
Em valiosas soluções.
No momento mais nublado em que tudo parecia perdido
Ele fez, na minha vida, o sol brilhar.
Hoje virtual na mais virtual idade.
Mundo magnífico onde a realidade é real
Onde o acaso esmera a ocasião.
Agradecido sempre
A quem me estendeu a mão.

Resumo

A presença de diferentes tipos e intensidades de ruídos nos sinais da fala, têm sido um desafio para definir um modelo para o reconhecimento automático da fala. Neste sentido, estuda-se a "reconstrução de características ausentes", que é um método de compensação, cujo objetivo é melhorar a robustez dos algoritmos de reconhecimento da fala em relação aos ruídos. Um modelo convencional para reconstrução de características ausentes utiliza características acústicas e métodos estatísticos para melhorar o reconhecimento da fala. No entanto, para este modelo, a taxa de acerto diminui quando o ruído presente no sinal é diferente do que foi utilizado no treinamento. Neste trabalho, um modelo adaptativo para reconhecimento da fala com reconstrução de características ausentes foi proposto. Para isso, foi utilizada uma nova abordagem para identificar as características articulatórias, através do pitch e do Mapa Auto-Organizável, e uma rede neural com topologia variante no tempo (LARFSOM) para reconstruir as características ausentes. O objetivo desse modelo é reconhecer a fala em sistemas *online* (tempo real) e *offline* que possam se modificar automaticamente sempre que for necessário. Assim, espera-se que o modelo seja independente de locutor. Para avaliar o modelo proposto, utilizamos as bases TIMIT e Aurora 2. Como resultados, foram obtidas uma taxa de erro médio de reconhecimento da fala de 6,96% para a base TIMIT e 4,46% para a base Aurora 2. Os experimentos realizados mostram que, mesmo sem utilizar um conhecimento prévio do sinal (oráculo), o modelo apresentou estabilidade (em relação a taxa de erro médio) quando existe presença ou ausência de ruído no sinal, bem como, na existência de locutores com diferentes gêneros e sotaques pronunciando frases com diferentes tamanhos.

Palavras-chave: Reconhecimento da Fala. Reconstrução de Características Ausentes. Características Acústicas e Articulatórias. Aprendizagem não-supervisionada. SOM. LARFSOM.

Abstract

The presence of different background noise in speech signal, has been a challenging to define a model for automatic speech recognition system. Missing-feature reconstruction is a compensation method to improve the noise robustness. A conventional models for missing-feature reconstruction is based on acoustic feature and statistical method to improve speech recognition. Nevertheless, these models degrade performance when different background noise is present in the signal. In this work, we propose a new adaptive speech model for speech recognition with missing-feature reconstruction, using unsupervised learning, for online (real-time) and offline systems, that automatically modifies as appropriate. For this, a new approach using Self-Organizing Map (SOM), to identify and extract articulatory features, and neural network with time-varying structure (LARFSOM), were used. In this work, an adaptive model for speech recognition with missing-feature reconstruction was proposed. For this, a new approach to identify the articulatory features, through the pitch and the Self-Organizing Map (SOM), and a neural network with time-varying structure (LARFSOM) for missing-feature reconstruction, were used. The purpose of this model is speech recognition in online (real-time) and offline systems, that automatically modifies as appropriate. Thus, it is expected that the model is robust for speaker variation. For evaluation purposes, Aurora 2 and TIMIT databases were used. As a result, we obtain a Word Error Rate average of 4.46% on Aurora 2 and 6.96% on TIMIT. Experimental results indicate that, even without prior knowledge (oracle) of the signal, the model is robust to noise, speaker variation, type of speech, and speech size.

Keywords: Robust Speech Recognition. Missing-feature Reconstruction. Articulatory and Acoustic Features. Unsupervised Learning. SOM. LARFSOM.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

ACF	Autocorrelation Function
AMDF	Average Magnitude Difference Function.
ASR	Automatic Speech Recognition.
CFR	Comb Filter Ratio.
DCT	Discrete Cosine Transform.
DDCT	Distributed Discrete Cosine Transform.
DRNN	Deep Recurrent Neural Networks.
EM	Expectation Maximization.
FFT	Fast Fourier Transform.
FIR	Finite Impulse Response.
GMM	Gaussian Mixture Models.
GNG	Growning Neural Gas.
GWR	Grow When Required.
HMM	Hidden Markov Model.
IIR	Infinite Impulse Response.
IPA	International Phonetic Alphabet.
LARFSOM	Local Adaptive Receptive Field Self-Organizing Map.
LHS	Latin Hypercube Sampling.
LIN	Linear Input Network.
LSTM	Long Short-Term Memory.
MFCC	Mel-Frequency Cepstral Coefficients.
MINERS	Model Invariant to Noise and Environment and Robust for Speech.
MIRS	Modified Intermediate Reference System.
MLE	Maximum Likelihood Estimates.
MLP	Multi-Layer Perceptron.
NCCF	Normalized Cross Correlation Function.
NIST	National Institute of Standards and Technology.
NMCCs	Normalized Modulation Cepstral Coefficient.

RASTA-PLP RelAtive SpecTrAl - Perceptual Linear Predictive.

RBM Restricted Boltzmann Machine.

RNN Recurrent Neural Network.

SFM Spectral Flatness Measure.

STFT Short Time Fourier Transform.

SUMÁRIO

1

Introdução

O som da voz é produzido pelo o ar, que é expelido dos pulmões, passa pela garganta, cordas vocais e sai pela boca. A fala é o som da voz articulada e moldada pela garganta, dentes, língua, bochechas e lábios. Quando se fala, pelo menos 12 músculos da laringe são movimentados e para cada fonema um conjunto de características (atributos), como por exemplo: energia, *pitch* e estrutura dos formantes, é produzido (LADEFOGED; JOHNSON, 2010).

O reconhecimento automático da fala, do inglês *Automatic Speech Recognition* (ASR), é um sistema que habilita a máquina a responder aos comandos, frase ou fala contínua pronunciada pelo locutor. Hoje, tal reconhecedor pode ser encontrado em dispositivos móveis, atendimento automático nos *call-centers*, dispositivos de autenticação, jogos eletrônicos, automação industrial, robótica, dentre outros (RABINER; SCHAFER, 2007).

As aplicações utilizando reconhecimento automático da fala, como por exemplo: reconhecimento de comandos e sintetizador da fala, tornaram-se uma das principais ferramentas adaptativas utilizadas por pessoas com deficiências visuais e motoras. Segundo o IBGE (2010) [1], cerca de 25,72% da população brasileira são portadores de deficiências visuais ou motoras, o que representa aproximadamente 49,65 milhões de pessoas. Esse cenário revela a importância de um aprimoramento nos reconhecedores da fala, proporcionando maior independência, qualidade de vida e inclusão social.

Um dos desafios em construir sistemas para reconhecer a fala é modelar o sinal que é obtido na produção da voz. Isso ocorre devido a quantidade de atributos da voz humana e das características próprias no contexto da fala (ruídos, sotaques e idiomas) (SINGH; STERN; RAJ, 2004).

Ruído é um som indistinto e sem harmonia, cuja intensidade é medida em decibéis (dB). A escala de decibéis é logarítmica, de modo que um aumento de três decibéis representa um aumento da intensidade de ruído para o dobro (RABINER; SCHAFER, 1978). Quando o ruído é inserido no sinal da fala, dificulta a extração das características (conhecida como descrição da fala) e, consequentemente, a correta classificação através das técnicas de reconhecimento, seja elas supervisionadas ou não-supervisionadas.

[1] http://www.ibge.gov.br/home/estatistica/populacao/censo2010/default.shtm, Visto em Março, 2017.

A reconstrução das caraterísticas ausentes da fala surgiu devido a degradação dos reconhecedores automáticos da fala, quando expostos a ambientes diferentes do seu treinamento. A diferença da reconstrução em relação ao reconhecimento é que na reconstrução é necessário identificar as características ausentes no sinal e depois repará-las, enquanto que no reconhecimento não há essa identificação. Os métodos de características ausentes (KIM; STERN, 2011), do inglês *missing-features*, são utilizados para reconstruir a fala. Estes métodos dependem das características da fala que foram menos afetadas pela ação do ruído. A princípio, os métodos deveriam ser capazes de melhorar o reconhecimento da fala, mesmo sem reconhecer a natureza do ruído, porém, a depender da intensidade e do tipo de ruído, esses métodos precisam passar por ajustes, perdendo, assim, o poder de generalização (RAJ; SELTZER; STERN, 2004).

Segundo KIM; STERN (2011), a reconstrução das características ausentes da fala pode ser dividida em duas etapas: a primeira é responsável por determinar uma máscara no plano tempo-frequência para identificar os componentes confiáveis e não confiáveis do sinal da fala. Esses componentes são trechos do sinal, geralmente representados pela energia ao longo do espectro, que sofreram alterações (componentes não confiáveis) ou que permanecem inalterados (componentes confiáveis) devido à ação do ruído. A segunda etapa é responsável por reconstruir os componentes que foram danificados pelo ruído (BARKER et al., 2000).

A dificuldade em reconstruir as características ausentes da fala passa pela definição de qual modelo utilizar. Sabendo que "modelo" é um conjunto constituído por um arcabouço e ferramentas necessárias para atingir um objetivo específico. Esses modelos variam a depender do nível de degradação do sinal, dificultando o uso em ambientes reais.

Para reconstruir as características ausentes da fala em tempo real é necessário lidar com a degradação do sinal em diversas intensidades. Essa necessidade ocorre em ambientes reais, pois as condições que o sistema foi treinado se diferem. VIANA; MELLO (2014), propuseram variações de descritores acústicos para lidar com amostras limpas e ruidosas, porém a intensidade e o local do ruído foram previamente definidos. Uma das soluções para resolver esse problema é a utilização de um método de aprendizagem que se adapte às condições do meio, sendo assim independente da variabilidade acústica.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo desta tese é propor um modelo adaptativo para o reconhecimento da fala com reconstrução de características ausentes, utilizando aprendizagem não-supervisionada, capaz de atuar em sistemas *online* (tempo real) e *offline*, que podem se modificar automaticamente sempre que for necessário.

1.1.2 Objetivo Específico

Como objetivos específicos temos:

1. Propor uma nova abordagem para extrair as características articulatórias;
2. Propor o uso de um rede neural com aprendizagem não-supervisionada e com topologia variante no tempo para reconstruir as características ausentes da fala;
3. Avaliar o desempenho, em relação a taxa de reconhecimento, do método "classificador dependente da frequência" sem utilizar o conhecimento prévio do sinal.

1.2 Breve Histórico dos Reconhecedores da Fala

A década de 1950 marcou o início do desenvolvimento de sistemas para o reconhecimento da fala. Esse interesse surgiu devido à evolução no campo da fonética e fonologia que exploravam as frequências fundamentais (*pitch*) e os formantes da fala. O*pitch* é uma característica acústica que fornece a percepção auditiva do tom (altura do som em relação a um ponto de referência). Os formantes são picos de energia, em uma região do espectro sonoro, ocasionado pelos vários componentes do trato vocal, possuindo um importante papel na inteligibilidade da fala. Em 1952, surgiu um dos primeiros sistemas automático para reconhecimento da fala que reconhecia dígitos de zero a nove de um único locutor. DAVIS; BIDDULPH; BALASHEK (1952) propuseram um circuito elétrico que realizava essa função. O reconhecedor usava os formantes da fala para identificar cada número pronunciado e tinha uma taxa de reconhecimento de 97%.

Essa abordagem não era capaz de reconhecer a fala de outro locutor. O circuito foi projetado para reconhecer a fala de um locutor em específico. Era necessário o tempo de pausa para cada dígito pronunciado, o locutor realizava uma pausa de 350ms antes de pronunciar o próximo número, e não reconhecia quaisquer palavras fora do alfabeto de dígitos.

Em 1959, o pesquisador FRY (1959) desenvolveu um sistema capaz de reconhecer quatro fonemas e nove consoantes da língua inglesa, através de um analisador de espectro e uma combinação de padrões. A técnica apresentou um baixo poder de generalização das palavras, dependência do locutor e não conseguia reconhecer dígitos (FRY, 1959).

Na década de 1960 surgiram os primeiros sistemas japoneses que reconheciam dígitos e fonemas. Autores como SUZUKI; NAKATA (1961) e NAGATA; KATO (1963) utilizaram regras de decisão de Bayes para reconhecer dígitos. Já SAKAY; DOSHITA (1962) empregaram a taxa de passagem pelo zero para identificar os padrões dos fonemas pronunciados. O sistema era dividido em três partes: classificador dos fonemas, circuito de controle e circuito de análise. As limitações do sistema era a incapacidade de reconhecer palavras e dependência de locutor, ou seja, o sistema só reconhecia as palavras para locutores específicos, os mesmos utilizados no treinamento e no teste.

Com os sistemas desenvolvidos na década de 1970, era possível reconhecer até 200 palavras com uma taxa de acerto de 97,3%. Como limitações, o sistema não reconhecia palavras contaminadas por ruído (sinal ruidoso), não existia a representatividade de todo o alfabeto e o sistema era dependente de locutor.

A década de 1980 marcou o início das pesquisas com fala contínua, quando não há definição prévia de pausas entre palavras. O modelo probabilístico, conhecido como modelo escondido de Markov, do inglês *Hidden Markov Model* (HMM), desenvolvido em 1966 (BAUM; PETRIE, 1966), tornou-se a principal ferramenta de classificação para a fala. O HMM é definido como um par de processos estocásticos que representam processos não observáveis, relacionados à variação temporal, e observáveis, relacionados à variabilidade espectral. Em geral, o HMM gera sequências de observações pulando de um estado para outro, emitindo uma observação a cada salto. O modelo oculto de Markov mais utilizado na área de reconhecimento da fala é o modelo *left-right*, ou modelo de Bakis (DELLER; HANSEN; PROAKIS, 2000), no qual a sequência de estados associada ao modelo tem a propriedade de, à medida que o tempo aumenta, o índice do estado aumenta ou permanece o mesmo. Isto é, os estados evoluem da esquerda para a direita no modelo.

Nos anos de 1990, a rede neural artificial *Multi-Layer Perceptron* (MLP) foi utilizada para fornecer probabilidade *a posteriori* para cada palavra em uma fala contínua. Para isso, um modelo híbrido utilizando HMM e MLP foi proposto (BOURLARD; MORGAN, 1994). Nesse modelo, a probabilidade *a posteriori* proveniente da MLP, foi utilizada como probabilidade de emissão para cada estado no modelo escondido de Markov.

A partir dos anos 2000, a hibridização das técnicas para o reconhecimento e reconstrução das características ausentes da fala ganhou notoriedade na área de processamentos de sinais. Com a evolução dos algoritmos de Bagging (BREIMAN, 1996), Boosting (MITCHELL, 1997) e AdaBoost (FREUND; SCHAPIRE, 1995), muitos pesquisadores começaram a utilizar os *ensembles* (máquinas de comitês) para classificar os sinais da fala (SAON; SOLTAU, 2012). Os *ensembles* são conjuntos de classificadores que se baseiam na ideia de unir as opiniões que os compõem para aumentar a precisão de um sistema de classificação de padrões. Cada classificador contribui com sua visão do espaço de características do problema apresentado, promovendo, assim, a diversidade entre seus integrantes.

1.3 Metodologia

O modelo proposto utiliza características acústicas e articulatórias. As características acústicas estão associadas as propriedades físicas dos sons da fala. Já as características articulatórias verificam como os sons são produzidos pelo aparelho fonador (LADEFOGED; JOHNSON, 2010). A utilização apenas das características acústicas dificulta o reconhecimento automático da fala quando há presença de ruído, pois, segundo LADEFOGED; JOHNSON (2010), essas características sofrem grandes variações em uma mesma palavra pronunciada pelo locutor em

diferentes ambientes, devido à não modelagem do efeito da coarticulação. Ou seja, há uma interferência de um segmento do sinal na realização de outro que, por exigências de ordem física, não há como os articuladores atingirem os mesmos alvos simultaneamente. Por isso, o uso das características articulatórias são importantes para o sistema de reconhecimento da fala, pois, tais características, não sofrem variações em uma mesma palavra pronunciada pelo locutor em diferentes ambientes.

Neste trabalho, a combinação das características acústicas e articulatórias e o mapa auto-organizável com campo receptivo adaptativo local, do inglês *Local Adaptive Receptive Field Self-Organizing Map* (LARFSOM), proposto por ARAUJO; COSTA (2007), foram utilizados. As características acústicas do sinal foram extraídas baseado em KIM; STERN (2011). Já as características articulatórias, um novo modelo utilizando *pitch* e a rede neural mapa auto-organizável, do inglês *Self-Organizing Maps* (SOM) (KOHONEN, 1997), foi proposto.

A rede LARFSOM foi utilizada para reconhecer a fala com reconstrução de características ausentes. A mesma incorpora características da rede SOM e o crescimento da rede quando necessário, do inglês *Grow When Required* (GWR) (MARSLAND; SHAPIRO; NEHMZOW, 2002).

Para avaliar o modelo proposto, foram utilizadas as bases TIMIT e Aurora 2 (PEARCE; HIRSCH; EUROLAB, 2000). Os resultados encontrados nos experimentos revelam uma taxa de erro médio de reconhecimento da fala de 6,96% para a base TIMIT e 4,46% para a base Aurora 2. Os experimentos realizados mostraram que, mesmo sem utilizar conhecimento prévio do sinal (oráculo), o modelo apresentou estabilidade (em relação a taxa de erro médio) quando há presença ou ausência de ruído no sinal. Bem como, quando há locutores com diferentes gêneros e sotaques pronunciando frases com diferentes tamanhos.

1.4 Estrutura do Documento

Além deste capítulo, esta tese é apresentada em mais cinco capítulos que estão organizados da seguinte forma:

Capítulo 2: São descritas as características acústicas e articulatórias utilizadas neste trabalho. Para isso, os conceitos básicos da fonética e fonologia são definidos para o melhor entendimento do trabalho. Será mostrada como a voz é produzida e será explicada sobre a variabilidade linguística.

Capítulo 3: O problema de reconstrução das características ausentes da fala é formalmente definido. Em seguida, trabalhos relevantes na reconstrução das características ausentes da fala são apresentados e analisados, possibilitando a discussão de suas capacidades e limitações.

Capítulo 4: É apresentado o modelo adaptativo para reconhecimento da fala com reconstrução de características ausentes, proposto neste trabalho.

Capítulo 5: São apresentados as metodologias utilizadas para realização dos experimentos. Bem como, a análise dos resultados.

Capítulo 6: São apresentados as contribuições para a ciência, a conclusão, os trabalhos publicados e os trabalhos futuros.

2

Características da Fala

A fonética e a fonologia são as áreas da linguística que estudam os sons da fala. A fonética estuda o ponto de vista articulatório, verificando como os sons são produzidos pelo aparelho fonador. Já a fonologia se dedica aos estudos dos sistemas de sons, a sua descrição, estrutura e funcionamento (LADEFOGED; JOHNSON, 2010).

Através da análise da fonética e fonologia, pode-se encontrar indicações presentes em sinais que possibilite a identificação de fonemas através da análise acústica e articulatória. Estas indicações são conhecidas como atributos da fala (RABINER; SCHAFER, 1978).

Este capítulo descreve as características articulatórias e acústicas da fala que foram utilizadas neste trabalho.

2.1 Produção da Voz

A voz é produzida a partir de três grupos de órgãos envolvidos: Sistema Respiratório, Sistema Fonatório e Sistema Articulatório. A Figura 2.1 apresenta os três sistemas citados. Portanto, de forma simples, o som da voz é produzido pelo o ar, que é expelido dos pulmões, passa pela garganta, cordas vocais e sai pela boca. A fala é o som da voz articulada e moldada pela garganta, dentes, língua, bochechas e lábios. Ao falar, o trato vocal muda de forma, movimentando pelo menos 12 músculos da laringe, produzindo diferentes sons (LADEFOGED; JOHNSON, 2010). O trato vocal é um tubo de ar fechado constituído pelo conjunto de órgãos responsáveis por produzir a fala. Alguns sons raros, como por exemplo, um clique na língua africana, são produzidos pela corrente de ar gerada por movimentos da laringe, enquanto a glote está fechada, não fazendo uso da corrente de ar da respiração. A Figura 2.2 mostra o trato vocal adaptada de (LADEFOGED; JOHNSON, 2010).

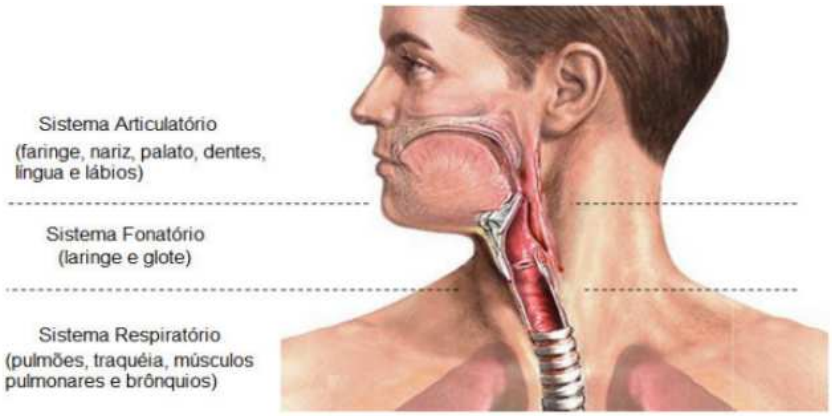


Figura 2.1: Os Sistemas: Respiratório, Fonatório e Articulatório. Adaptada de (BISOL, 2001).

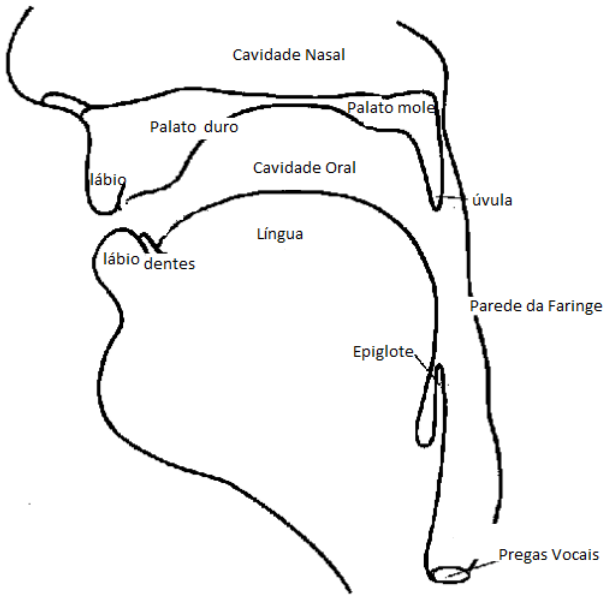


Figura 2.2: Trato vocal. Adaptada de (LADEFOGED; JOHNSON, 2010).

A perfeita sincronia desses três grupos de órgãos, possibilita a produção de uma voz entendível denominada de fala.

O conjunto limitado de sons que os humanos conseguem produzir é classificado em quatro tipos: os sons vozeados ou sonoros que representam o vibrar das cordas; os sons surdos ou não vozeados, onde as cordas vocais não vibram, apenas permanecem abertas; os sons explosivos, que resultam do fechamento completo do trato vocal e os sons de excitação mista, que combinam a vibração das pregas vocais, sons sonoros, com a excitação não vozeada, sons surdos (LADEFOGED; JOHNSON, 2010).

2.2 Características Articulatórias

Com intuito de explorar os métodos para descrição, classificação e transcrição dos sons da fala. A fonética, segundo LADEFOGED; JOHNSON (2010), se divide em três focos de estudos:

- Fonética Articulatória: Descreve como a fala é produzida do ponto de vista articulatório e fisiológico.
- Fonética Auditiva: Compreende o estudo da percepção da fala.
- Fonética Acústica: Compreende o estudo das propriedades físicas dos sons da fala, a partir da sua transmissão do falante ao ouvinte.

A presença ou ausência de obstrução na passagem de ar pela cavidade supraglotal (cavidade que engloba as partes oral, nasal e a faringe) produz sons que são classificados como: glides ou semivogais, vogais e consoantes. Caso o ar sofra obstrução o som é classificado como consoantes, caso contrário, é classificado como vogal. Entretanto, existem aqueles sons que a passagem do ar não é definida, sendo classificados como glide ou semivogais (BISOL; BRESCANCINI, 2002).

Cada vogal ou consoante é diferenciada pela forma articulatória que é produzida. Por conta disso, a Associação Fonética Internacional criou uma classificação desses segmentos conhecida como Alfabeto Fonético Internacional, do ingês *International Phonetic Alphabet* (IPA). Nos quais vogais, consoantes e segmentos que não se enquadram em nenhum dos dois, são classificados de acordo a forma de articulação. A principal diferença entre a articulação das vogais e das consoantes está no fato de que, para identificar a vogal, precisa-se olhar a totalidade da cavidade oral, pois há uma ausência de obstrução à passagem do ar pela boca.

Para emitir uma vogal, o ápice da língua se desloca no interior do aparelho fonador, tanto no eixo horizontal quanto no eixo vertical. Deslocando-se na horizontal, a língua vem para frente ou recua para o fundo da boca. Ao deslocar-se na vertical, a língua sobe ou desce. Todo esse deslocamento lembra um trapézio com a base menor para baixo. Os foneticistas chamam esse processo de deslocamento de trapézio vocálico (HORA; COLLISCHONN, 2003). A Figura 2.3 mostra o trapézio vocálico.

As consoantes são classificadas de acordo com a maneira e o lugar de articulação. Em LADEFOGED; JOHNSON (2010), a maneira de articulação das consoantes é classificado como:

- Oclusivas: O som é produzido por um bloqueio na corrente de ar. Ex.: **p**ato;
- Nasais: O som é produzido com o bloqueio do ar na cavidade oral e o rebaixamento do palatino, o qual permite a passagem de ar pelas narinas. Ex.: da**m**a;
- Fricativos: O som é produzido com o estreitamento de alguma parte do aparelho fonador, sofrendo fricção. Ex.: **f**aca;

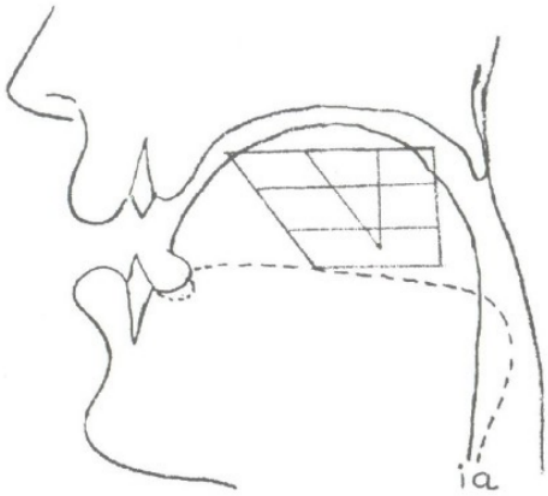


Figura 2.3: Trapézio vocálico. Adaptada de (HORA; COLLISCHONN, 2003).

- Africados: O som é produzido inicialmente pelo bloqueio da passagem de ar dentro da cavidade oral, sofrendo posteriormente uma obstrução que provoca fricção. Ex.: **T**iago;
- Laterais: A cavidade oral anterior bloqueia a passagem central do ar, permitindo apenas uma passagem lateral. Ex.: **l**abirinto, ca**lh**a;
- Vibrantes ou vibrantes múltiplos: Caracterizados por várias batidas rápidas da língua no véu palatino (de múltiplos contatos). Ex.: ca**rr**o;
- Vibrante simples ou tepe: Uma batida rápida da ponta da língua nos alvéolos dos incisos superiores, provocando uma rápida obstrução do ar. Ex.: b**ra**vo;
- Retroflexo: O som é produzido pelo curvamento da ponta da língua para cima e para trás, como na pronúncia do "r" nos sotaques do interior de alguns estados como São Paulo. Ex.: ca**r**ne;
- Aproximantes: São sons formados acima da área das vogais, mas a passagem de ar é maior que a pressão que causa a fricção. Ex.: **l**ábios;

No que diz respeito ao lugar da articulação, em MATEUS; FALÉ; FREITAS (2005) encontra-se:

- Bilabial: Essa consoante é formada pela obstrução da passagem do ar que resulta no movimento de um lábio contra o outro, sendo que o lábio inferior é o articulador ativo e o lábio superior é o articulador passivo. Ex.: **p**ato, **m**ãe;
- Labiodental: O articulador ativo é o lábio inferior e o passivo são os dentes incisivos superiores. Ex.: **f**aca, **v**aca;
- Dental: Nessa consoante, o articulador ativo é a língua (ápice ou lâmina), e seus articuladores passivos são os dentes incisivos superiores. Ex.: **d**ata;

- Alveolar: São as consoantes cujo som é articulado no encontro da ponta da língua com os alvéolos dentários. O articulador ativo é a língua (ápice ou lâmina) e o passivo são os alvéolos. Ex.: **l**ata;
- Palatoalveolar: É produzido na região imediatamente posterior à região onde o som alveolar é produzido. Ex.: **ch**umbo;
- Alveolopalatal: Essas consoantes também são chamadas de pós-velares. Onde o articulador ativo é a parte anterior da língua e o passivo é a parte medial do palato duro (céu da boca). Ex.: **t**ia, **d**ia;
- Palatal: A sua pronúncia é formada pela aproximação ou o contato do dorso da língua com o palato duro. O articulador ativo é a parte média da língua e o passivo é a parte final do palato duro. Ex.: **p**alha;
- Velar: É formado pela aproximação ou o contato da língua com o palato mole (véu palatino). O articulador ativo é a parte posterior da língua e o passivo é o palato mole. Ex.: **g**ata, **r**ata;
- Uvular: É produzida pela parte posterior da língua pressionando o fundo da cavidade oral (palato mole e úvula);
- Faringal: É produzida pela constrição da ponta da língua com a faringe. Na língua portuguesa, não há palavras com essa constrição. Há muitas ocorrências da faringal na língua árabe.
- Glotal: Em sua pronúncia, o ponto de articulação é a glote que se comporta como articuladores. Ex.: esca**rrar**, pronunciando o /r/ ao mesmo tempo.

Na Figura 2.4 é mostrada a classificação das vogais e consoante adaptada da tabela IPA[1].

As características articulatórias podem ser obtidas por três formas: a) Através de uso de equipamentos para medir diretamente o trato vocal, como por exemplo, através da cinerradiografia (PAPCUN et al., 1992); b) Através das características acústicas, utilizando a inversão dos filtros (RICHARDS et al., 1997), utilizada para encontrar variação da trajetória da fala; c) Com a combinação das probabilidades de cada classe fornecidas pelas redes (KIRCHHOFF; FINK; SAGERER, 2002). A terceira forma de se extrair características articulatórias são chamadas de pseudo-articulatórias, uma vez que não fornecem detalhes sobre o trato vocal. Diversos trabalhos utilizaram características articulatórias para descrever a fala, dentre eles destacam-se:

1. DENG; YU (2014) utilizaram características articulatórias de maneira e lugar de articulação para ASR. Os autores utilizaram o HMM para reconhecer os fonemas da base TIMIT. Como resultado, obtiveram uma taxa média de acerto de 72%. Esse resultado representa uma superioridade de 15%, em relação a taxa de reconhecimento, quando comparado ao uso do descritor coeficientes cepstrais de frequência mel, do inglês *Mel-Frequency Cepstral Coefficient* (MFCC).

[1] http://www.langsci.ucl.ac.uk/ipa/, Visto em Março, 2017.

Consoantes

⇩ Maneira / Lugar ⇨	Bilabial	Labiodental	Dental	Alveolar	PalatoAlveolar	Retroflexo	Palatal	Velar	Uvular	Faringal	Glotal
Oclusiva	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Vibrantes Simples	ʙ	ⱱ		r					ʀ		
Vibrantes Múltiplos				ɾ		ɽ					
Fricativa	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Laterais				ɬ ɮ							
Aproximante		ʋ		ɹ		ɻ	j	ɰ			
Retroflexo				ɭ		ɭ̆	ʎ	ʟ			

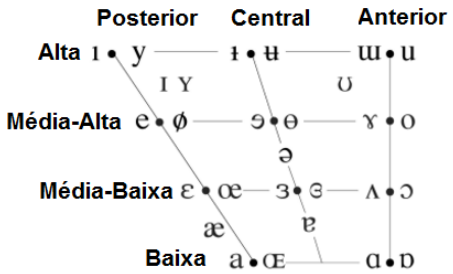
(a)

Africada

t͡s	Alveolar Africada Surda
t͡ʃ	Palatoalveolar Africada Surda
t͡ɕ	Alveolopalatal Africada Surda
t͡ʂ	Retroflexa Africada Surda
d͡z	Alveolar Africada Sonora
d͡ʒ	Palatoalveolar Africada Sonora
d͡ʑ	Alveolopalatal Africada Sonora
d͡ʐ	Retroflexa Africada Sonora

(b)

Vogal



(c)

Figura 2.4: Classificação das vogais adaptada da tabela *IPA*[1]: (a) Tabela das Consoantes, (b) Tabela da Consoante Africada, (c) Classificação das Vogais.

2. KIRCHHOFF; FINK; SAGERER (2002) foram um dos primeiros trabalhos a estimar as características articulatórias, utilizando a combinação das redes neurais com HMM, para reconhecimento automático da fala. Os autores utilizaram características pseudo-articulatórias e compararam o uso dessas características aplicadas ao ambiente com e sem presença de ruídos, em relação às características acústicas. Como resultado, os autores obtiveram uma taxa de acerto similar entre as características articulatórias (91,6%) e acústicas (91,1%) em relação a sinais sem ruído. Porém, na presença de ruído, em média, o uso das características articulatórias teve uma taxa de acerto (71,8%) melhor do que o uso das características acústicas (69,27%) para sistemas de reconhecimento automático da fala.

3. DHANANJAYA; YEGNANARAYANA; SURYAKANTH (2011) utilizaram informações acústicas da excitação do trato vocal, chamado de *epoch*, para corrigir os erros de reconhecimento de fonemas na base TIMIT. Para isso, foram utilizadas características articulatórias da maneira de articulação da fala (presença/ausência da fala, fricativo e nasal) e HMM. Como resultados, os autores tiveram uma taxa de acerto de 95,3% na identificação da presença da fala (fonemas).

4. NAESS; LIVESCU; PRABHAVALKAR (2011) compararam o algoritmo K-Vizinhos mais próximos, do inglês *K-Nearest Neighbor*, com a rede perceptron de múltiplas

camadas, do inglês *Multi-Layer Perceptron* (MLP), para identificar as características articulatórias (a partir da trajetória articulatória do trato vocal). Para isso, subpalavras foram utilizadas em vez de fonemas. Como resultado, o KNN apresentou uma melhora entre 2% a 6,4%, em relação à taxa de reconhecimento, comparado com a rede MLP (40% de acerto).

5. MITRA et al. (2013) utilizaram trajetórias articulatórias, como características, para sinais limpos e ruidosos, aplicadas aos sistemas de reconhecimento automático da fala. Para isso, os autores compararam as trajetórias provenientes dos descritores acústicos MFCC, *Relative Spectral - Perceptual Linear Predictive* (RASTA-PLP) (HERMANSKY; MORGAN, 1994) e normalização modular dos coeficientes cepstrais, do inglês *Normalized Modulation Cepstral Coefficients* (NMCCs) (MITRA et al., 2012). Como treinamento, os autores utilizaram sinais sem ruído fornecido pela base Aurora 4 e testaram utilizando os sinais ruidosos da mesma base. Como resultado, o NMCCs apresentou melhores resultados quando testados com os mesmos canais para o treinamento, taxa de erro de 36,5%, e quando testados com canais diferentes do treinamento, taxa de erro de 42,2%, em relação aos outros descritores.

6. MANJUNATH; RAO (2016) recentemente exploraram uma base Indiana com intuito de extrair boas características articulatórias em locuções previamente estabelecidas, quando existem palavras escritas e o locutor apenas faz a leitura, locuções de improviso, quando não há palavras escritas para a leitura, e conversação, onde há duas ou mais pessoas conversando sem um discurso pré-estabelecido. Para isso, os autores utilizaram MFCC, com intuito de estimar as características articulatórias, redes neurais e HMM para classificar as características. Como resultados, os autores obtiveram em média uma taxa de acerto (reconhecimento) de 73% para locuções previamente estabelecidas, 63% para locuções de improviso e 64% para conversação.

2.3 Características Acústicas

As características acústicas fornecem informações sobre o trato vocal. Os atributos acústicos da fala são: frequência fundamental (*pitch*); energia; estrutura dos formantes; número de picos e taxa de cruzamento por zero. A frequência fundamental é determinada pelo número de vibração das cordas vocais. Nos homens a frequência fica entre 80 a 150 Hz e nas mulheres entre 150 a 250 Hz (BEHLAU, 1995).

A energia tem o papel de medir a intensidade sonora. É através dela que é feita a diferenciação entre segmentos surdos e sonoros do sinal de voz. Isso ocorre devido à amplitude nos segmentos surdos ser mais baixa do que nos segmentos sonoros. Portanto, para medir a energia, técnicas no domínio do tempo (análise temporal) ou no domínio da frequência (análise espectral) são utilizadas (RABINER; SCHAFER, 1978).

A estrutura dos formantes fornecem indicações de como os fonemas são formados. É através dela que é identificada a duração da fala, que depende da velocidade com que os fonemas são pronunciados, a pausa e a entonação (variação da frequência fundamental) (RABINER; SCHAFER, 1978).

Por sua vez, o número de picos do sinal auxilia na identificação dos sons fricativos sonoros. Exemplo, o /f/, que pode ser confundido com vogais de pequena intensidade. Por outro lado, a taxa de cruzamento por zero auxilia nas identificações dos sons surdos, por exemplo, o /s/ (RABINER; SCHAFER, 1978).

Os atributos acústicos são os mais utilizados nos sistemas de reconhecimento da fala. Eles são responsáveis por diferenciar as palavras e são extraídos a partir de descritores como: MFCC e o coeficiente cepstral potencialmente normalizado, do inglês *Power-Normalized Cepstral Coefficient*(PNCC).

Devido à variabilidade do microfone e ambiente, os descritores da voz podem apresentar dificuldades em representar os atributos da fala. Esta dificuldade se dá devido à presença do ruído; variações de distância entre a boca e o microfone; velocidade da pronúncia; variação dos períodos de pausa; dentre outros. Para minimizar os efeitos, faz-se necessária a utilização de um pré-processamento no sinal da voz, com intuito de deixar o sinal mais próximo da fala "limpa".

Uma das etapas do pré-processamento é a utilização do janelamento na voz. Esse pré-processamento é necessário devido à natureza da variação do sinal da fala. Sendo comum dividir o sinal em janelas ou molduras, do inglês *frames*, realizada segundo os princípios da análise em curto prazo, dada pela Equação 2.1 (RABINER; SCHAFER, 2007).

$$X_n = \sum_{m=-\infty}^{\infty} x[m]w[n-m] \tag{2.1}$$

onde: X_n são os parâmetros em um tempo de análise n; $w[n-m]$ corresponde a uma janela de observação centrada em n ao longo do tempo; $x[m]$ é o sinal em análise.

O janelamento minimiza as margens de transição em formas de onda truncadas, reduzindo a perda espectral. A janela de Hamming (SONG; PENG, 2008) é a mais utilizada para esta função, que pode ser matematicamente representada pela Equação 2.2.

$$w[n] = \begin{cases} 0,54 - 0,46\cos(\frac{2\pi n}{N}), & n = 0,1,...,N\text{-}1 \\ 0, & \text{caso contrário} \end{cases} \tag{2.2}$$

onde: $w[n]$ é o sinal em análise e N é o tamanho da janela.

O tamanho de cada janela e da superposição é escolhido de acordo com o experimento proposto. A Figura 2.5 mostra a representação em tempo discreto da janela de Hamming. Na Figura 2.6 tem-se uma visão geral da superposição das janelas de Hamming aplicadas a um sinal.

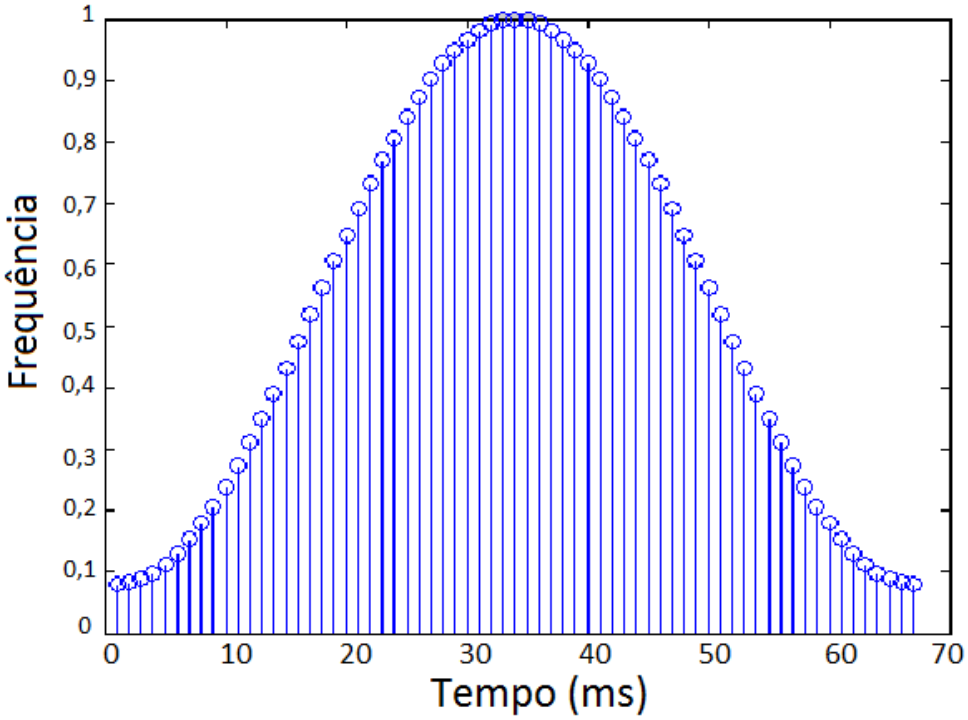


Figura 2.5: Janelas de Hamming aplicadas a um sinal. Adaptada de SONG; PENG (2008).

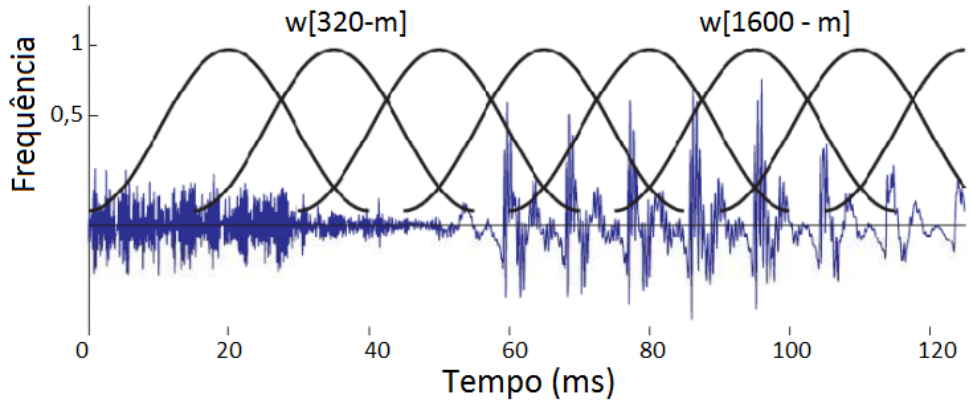


Figura 2.6: Superposição das janelas de Hamming. Adaptada de SONG; PENG (2008).

Jean-Baptiste Joseph Fourier propôs uma transformada que leva seu nome, transformada de Fourier. Essa transformada permite que funções não periódicas, mas cuja área sob a curva é finita, sejam expressas como uma integral de senos e/ou cossenos multiplicada por uma função de ponderação (POLUR; MILLER, 2005). A Equação 2.3, mostra como obter a transformada de Fourier de uma função contínua $f(t)$ de uma variável contínua, t, expressa por $F(\omega)$.

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \tag{2.3}$$

onde:

$\omega = 2\pi\mu$;
$\mu = variável\ contínua$;
$j = \sqrt{-1}$;
$e^{-j\omega t} = \cos(\omega) - jsen(\omega)$.

A transformada inversa de Fourier é realizada para conseguir obter o sinal original após a execução da transformada Fourier:

$$f(t) = F^{-1}(F(\omega)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}dt \tag{2.4}$$

Nas Figuras 2.7 e 2.8 podem ser observadas a aplicação da transformada de Fourier no sinal da frase "The birch canoe slid on the smooth planks" gravada em ambiente sem ruído e com ruído a 0dB, respectivamente. Pode-se observar uma maior intensidade no sinal na Figura 2.8, ocasionado pelo ruído do ambiente.

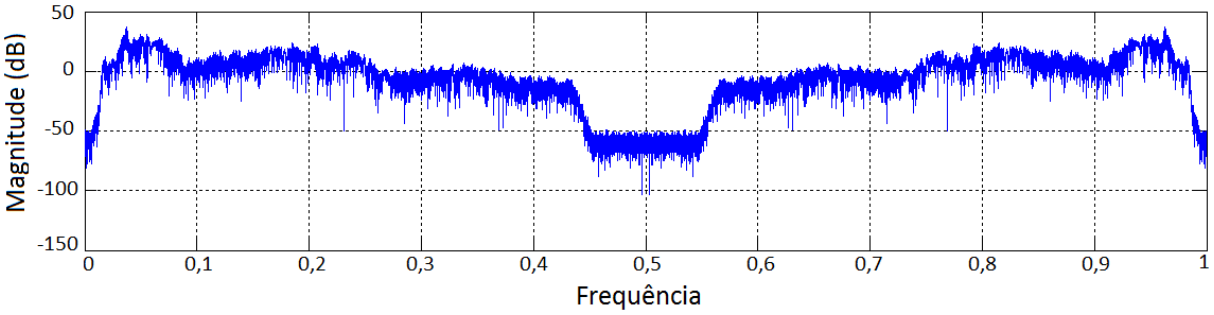


Figura 2.7: Transformada de Fourier aplicada a frase "The birch canoe slid on the smooth planks".

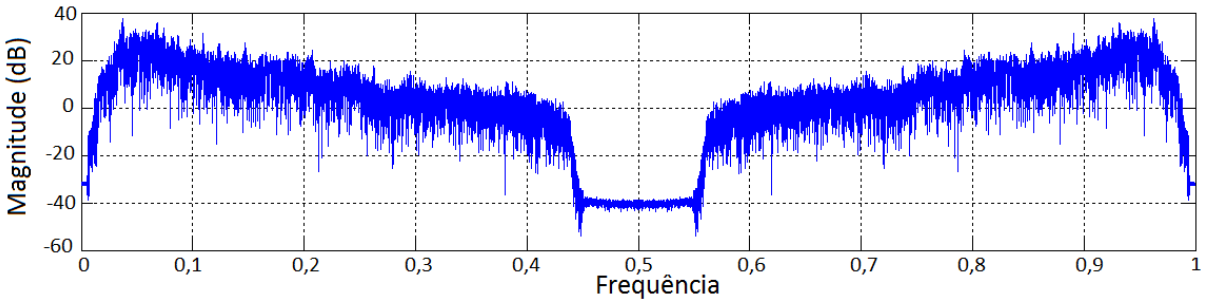


Figura 2.8: Transformada de Fourier aplicada a frase "The birch canoe slid on the smooth planks" com presença de ruído a 0dB.

Apesar da transformada de Fourier ser muito utilizada nos descritores da voz, ela permite apenas a análise de características no domínio da frequência, não possibilitando a completa determinação da relação espaço frequência, ou seja, a transformada é capaz de revelar quais frequências estão no sinal, mas não onde elas se encontram (KUHNE; TOGNERI; NORDHOLM, 2007). Baseado nesta dificuldade, Dennis Gabor adaptou a transformada de Fourier para ser

aplicada em pequenas janelas, denominada de janelamento do sinal, do inglês *windowing the signal*. Esta adaptação ficou conhecida como transformada de Fourier de tempo curto, do inglês *Short Time Fourier Transform* (STFT) (RABINER; JUANG, 1993).

A STFT mostra informações entre o tempo e a frequência do sinal, sendo possível identificar quando e em que frequência o evento de um sinal ocorreu. A STFT é a mais aplicada nos estudos de reconhecimento da fala. A desvantagem da técnica é a incapacidade de redimensionar o tamanho da janela ao longo do sinal, isto é, quando definido o tamanho da janela ela será a mesma ao longo do sinal. As Figuras 2.9 e 2.10 mostram o espectrograma gerado pela STFT da frase "*The birch canoe slid on the smooth planks*" com ausência e presença de ruído a 0dB gravado em um aeroporto, respectivamente. O tons avermelhados no espectrograma revela à intensidade sonora da amostra. Pode-se observar que na Figura 2.9 (amostra sem ruído) a intensidade sonora é menor que na Figura 2.10 (amostra com ruído).

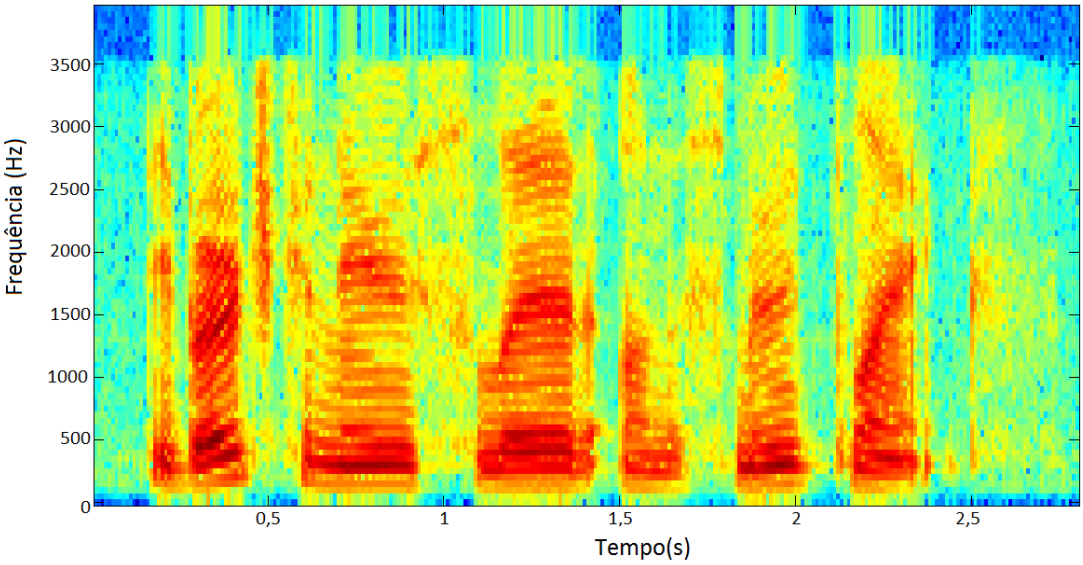


Figura 2.9: Espectrograma da STFT para a frase "*The birch canoe slid on the smooth planks*".

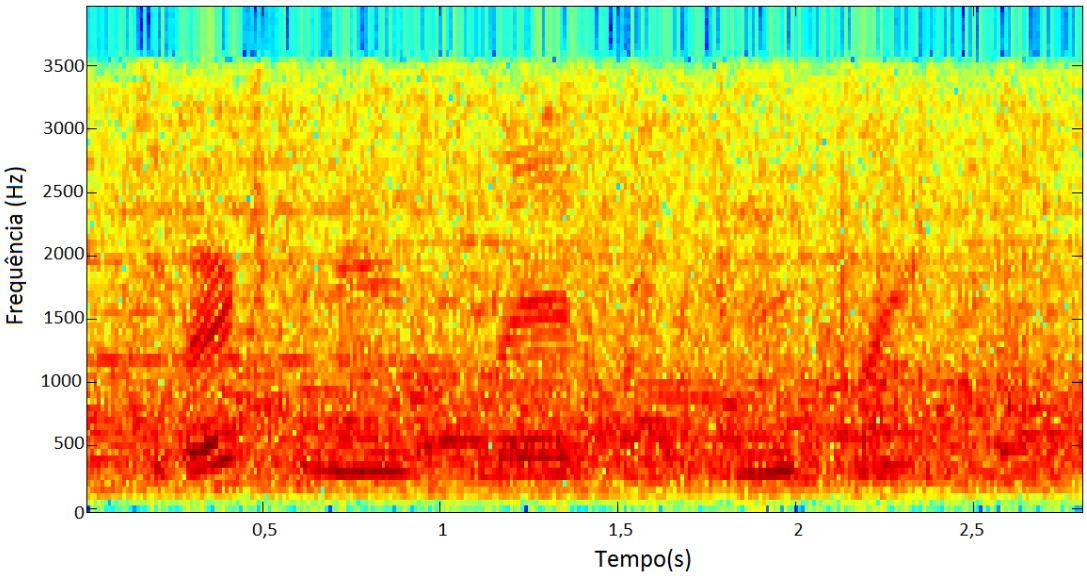


Figura 2.10: Espectrograma da STFT para a frase "*The birch canoe slid on the smooth planks*" com presença de ruído a 0dB.

2.3.1 MFCC

Os coeficientes *Mel-Cepstrais* surgiram devido aos estudos na área de psicoacústica (ciência que estuda a percepção auditiva humana), os quais revelam que a percepção humana das frequências de tons puros não segue uma escala linear. Através dessa análise, surgiu a ideia de serem definidas frequências subjetivas de tons puros. Para cada tom com frequência f, medida em Hz, define-se um tom subjetivo medido em uma escala psicoacústica que se chama escala mel (RABINER; JUANG, 1993).

A Mel é uma unidade de medida da frequência percebida de um tom. Como referência, definiu-se a frequência de 1kHz, com potência 40dB acima do limiar mínimo de audição do ouvido humano, como 1000 mels. Os outros valores foram obtidos através de experimentos, onde foi observado que a escala em Hz e a escala em Mel são aproximadamente linear abaixo dos 1000Hz e logarítmica acima deles. Logo, a escala Mel faz com que as faixas de frequência sejam posicionadas em uma escala logarítmica, a qual se aproxima da resposta do sistema auditivo humano (STEVENS; NEWMAN, 1937).

As equações que fazem a conversão da escala Mel para Hz e Hz para Mel são mostradas nas Equações 2.5 e 2.6, respectivamente. A Figura 2.11 mostra a relação entre a escala Mel e a escala Hz, adaptada de STEVENS; NEWMAN (1937).

$$M = 1127,01048 \log_e \left(1 + \frac{f}{700}\right) \tag{2.5}$$

$$f = 700\left(e^{\frac{M}{1127,01048}} - 1\right) \tag{2.6}$$

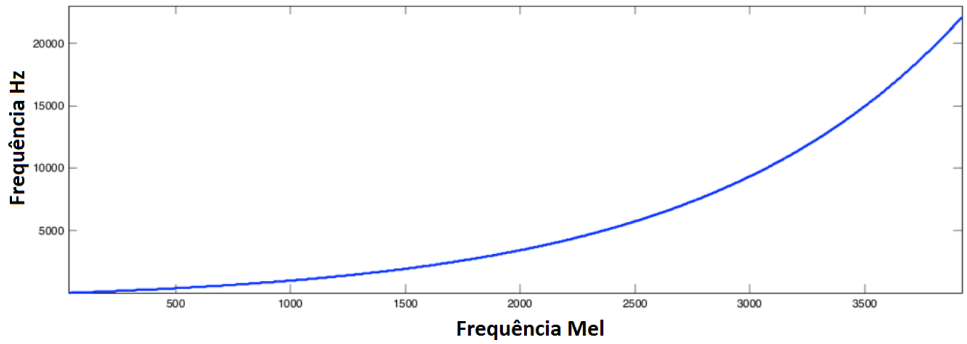


Figura 2.11: Relação entre a escala Mel e a escala Hz. Adaptada de STEVENS; NEWMAN (1937).

Além da escala Mel para definir os coeficientes do MFCC, aplica-se a transformada rápida de Fourier, do inglês *Fast Fourier Transform* (FFT), o banco de filtro triangular espaçado pela escala Mel e a transformada discreta do cosseno, do inglês *Discrete Cosine Transform* (DCT), (PATEL; RAO, 2010). A Figura 2.12 mostra o diagrama para o cálculo do MFCC.

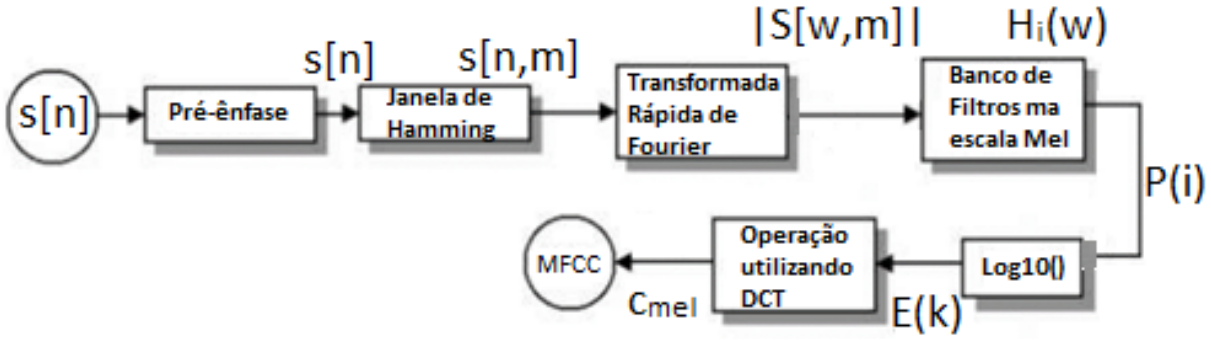


Figura 2.12: Diagrama para o cálculo do MFCC. Adaptada de PATEL; RAO (2010).

Inicialmente, é utilizada a pré-ênfase no sinal da voz, para compensar a atenuação dos componentes de alta frequência, causadas pelo mecanismo da produção da voz. Após essa etapa, divide-se o sinal da voz em janelas utilizando Hamming. Para cada trecho do sinal obtido, calcula-se a transformada rápida de Fourier.

A maior utilidade da escala Mel está na criação do banco de filtro, constituído por superposição de filtros triangulares. Estes filtros possuem frequências centrais linearmente espaçadas e a largura de banda é espaçada conforme a escala Mel. Para a fala humana são utilizados entre 12 a 40 filtros, escolhidos experimentalmente. A Figura 2.13 foi gerada com o auxílio do software MatLab [2] e mostra o banco de filtros triangulares composto por 20 filtros;

[2]Criado pela MathWorks Inc., o MatLab é um software que permite: a manipulação de matrizes, a criação de gráficos de funções e de dados, a criação e execução de algoritmos, além de possuir uma vasta gama de funções pré-definidas.

frequência do sinal da fala de 8000 Hz e duração de 256 ms para cada janela.

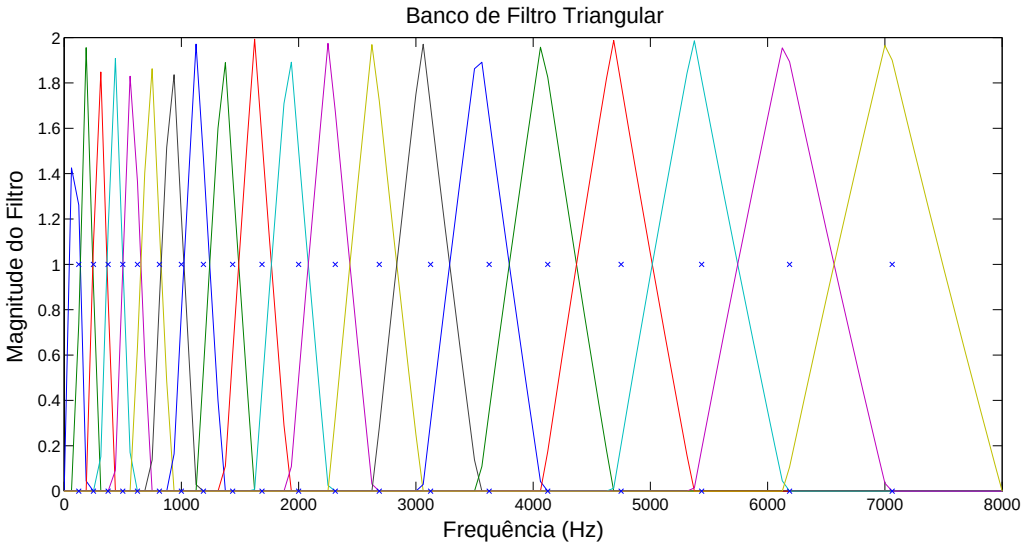


Figura 2.13: Banco de Filtro Triangular.

A última etapa para obter os coeficientes do MFCC é através do uso da DCT (BLINN, 1993). Esta técnica é utilizada para a compressão dos dados, fazendo uso apenas de números reais. Como resultado, é possível ver o acúmulo dos coeficientes mais significativos no início do vetor, excluindo os valores com pouca ou nenhuma informação. A Equação 2.7 mostra o cálculo da DCT.

$$X[k] = \sum_{n=0}^{N-1} x[n] \cos[\frac{\pi}{n}(n+\frac{1}{2})k] \tag{2.7}$$

onde: *X(k)* são os coeficientes resultantes da transformada discreta do cosseno; *x(n)* o sinal da fala; *N* o número de coeficientes.

Segundo PATEL; RAO (2010), de modo simplificado, podem-se obter os coeficientes do MFCC através da seguinte equação:

$$c(n) = \sum_{k=1}^{M} \log_{10} X(k) \cos(N(\frac{k-1}{2})\frac{\pi}{M}) \tag{2.8}$$

onde: $1 \leq n \leq N$; *X(k)* é a energia na saída do k-ésimo filtro; *M* é o número de filtros e *N* é o número de coeficientes.

Trabalhos como AMITA; BANSAL (2010) e HOSSAN; MEMON; GREGORY (2010) mostram o descritor MFCC aplicado à amostra ruidosa e sem ruído. O primeiro autor aplica o MFCC em uma base indiana, enquanto que o segundo autor modifica a etapa DCT do descritor

MFCC, propondo a utilização da técnica chamada de transformada discreta dos cossenos distribuída, do inglês *Distributed Discrete Cosine Transform* (DDCT). No entanto, os autores revelam em seus experimentos o baixo poder de descrição do MFCC quando exposto a amostra ruidosa. Devido a essa dificuldade, novas técnicas foram propostas para descrever a voz com intuito de aumentar a taxa de reconhecimento da fala em ambientes ruidosos. Dentre elas destaca-se o MINERS.

2.3.2 MINERS

O modelo robusto para fala invariante ao ruído e ambiente, do inglês *Model Invariant to Noise and Environment and Robust for Speech* (MINERS), foi proposto por VIANA; MELLO (2014) com objetivo de extrair características acústicas da voz, independente do ambiente e da presença ou ausência do ruído. Esse descritor foi desenvolvido seguindo as etapas:

1. Classificação do sinal como ruidoso ou não;
2. Utilização da Transformada *Wavelet* (MORLET et al., 1982) combinada com o PNCC2;
3. Utilização da técnica MFCC.

A Figura 2.14 mostra o processo decisório do descritor MINERS. Já a Figura 2.15 mostra as etapas da combinação do *Wavelet* com PNCC2. O PNCC2 foi assim denominado por VIANA; MELLO (2014) como um descritor PNCC com mascaramento temporal.

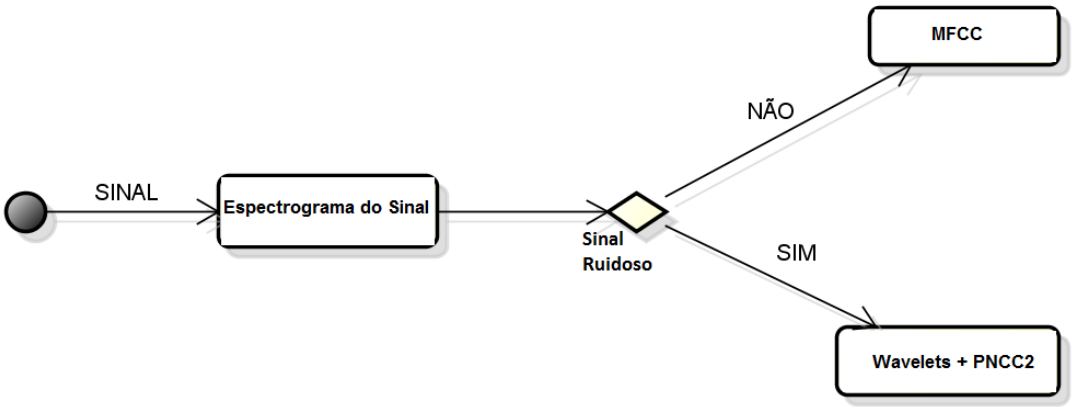


Figura 2.14: Fluxograma do descritor MINERS.

Para classificar um sinal como ruidoso ou não, os autores utilizaram a imagem da representação do sinal da fala. Por exemplo, considere um sinal da fala ao qual é adicionado um ruído de 5dB. Na Figura 2.16.a, é mostrada a transformada de Fourier do sinal limpo e, na Figura 2.16.b, tem-se a transformada do sinal com ruído.

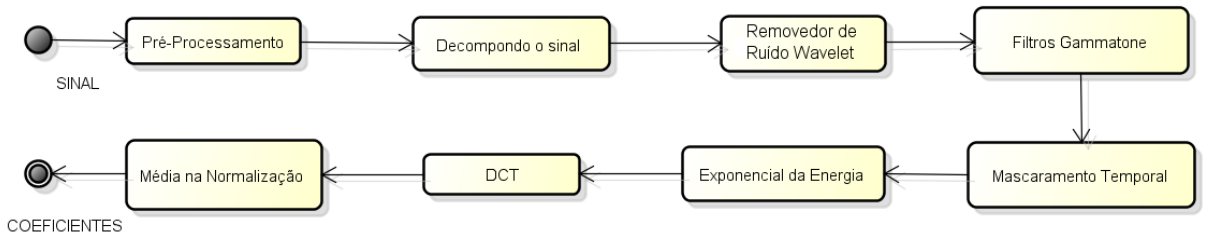


Figura 2.15: Wavelet combinado com PNCC2.

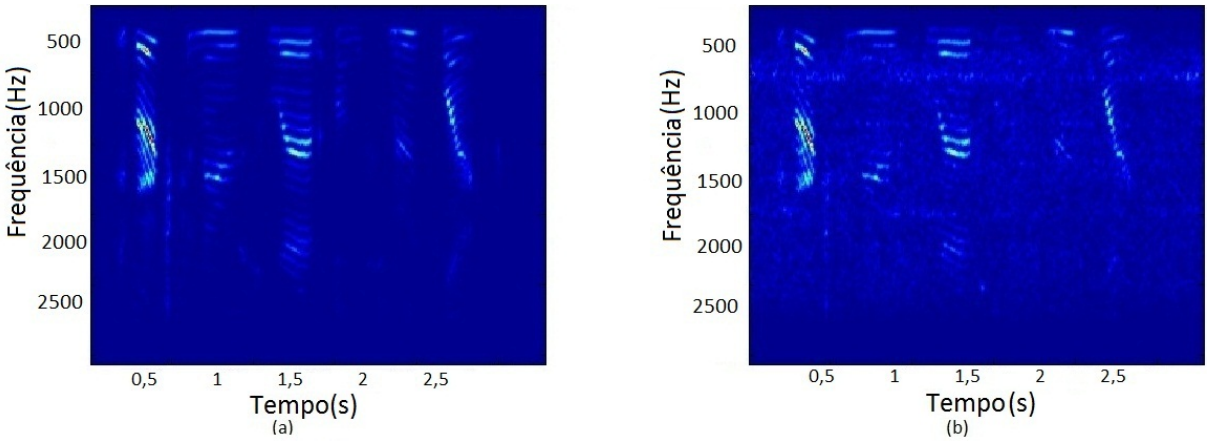


Figura 2.16: Transformada de Fourier de um sinal (a) sem ruído e (b) com ruído a 5dB.

Na representação gráfica, uma imagem é criada onde cada tom está associado à amplitude do gráfico naquela coordenada. Tons azulados indicam baixas amplitudes, enquanto tons avermelhados indicam as mais altas amplitudes.

Para classificar o sinal como ruidoso ou não, os autores localizaram na imagem dos coeficientes MFCC, os tons com valores de vermelho maiores que 120 e de verde menores que 130. Após a identificação das imagens, uma operação morfológica de fechamento, com elemento estruturante na forma de um disco de raio 2 (GONZALEZ; WOODS, 2008), foi aplicada à imagem gerada pela complementação dessas imagens (ou seja, os tons pretos são convertidos para branco e vice-versa). A Figura 2.17 mostra o resultado dessa operação (após uma nova complementação).

Figura 2.17: Resultado da aplicação de uma operação de fechamento morfológico nas imagens das Figuras (a) 2.16.a e (b) 2.16.b.

A partir dessa imagem, é calculada a percentagem de tons pretos. Imagem que tiver menos de 22% de tons pretos, é considerada sem ruído, caso contrário, mais de 78% de tons pretos, são consideradas com ruído. Esses parâmetros foram definidos experimentalmente. Quando o sinal é classificado como "sem ruído", o sinal é descrito pelo MFCC. Caso contrário, o sinal é descrito pela combinação das técnicas *Wavelet* e PNCC2.

Essa combinação foi proposta através da análise do comportamento de cada técnica, com intuito de lidar com amostras em diversos níveis de ruído e ambiente. Para isso a saída da transformada *Wavelet* (MALLAT, 2009) do tipo Daubechies (DAUBECHIES, 1992) é passada como entrada para o PNCC2.

3

Reconstrução das Características Ausentes Fala

A degradação dos reconhecedores automático da fala ocorre por diversos motivos. Dentre eles, está a ação do ruído ou quando o sinal da fala é exposto a um ambiente diferente do seu treinamento (GONZALEZ et al., 2013).

Várias técnicas foram propostas para diminuir a degradação do sinal ocasionada pelo ruído. Essas técnicas são denominadas de métodos de compensação. ACERO; ALEJANDRO (1992), BOLL (1979), GALES; YOUNG (1996) e LAMEL; KASSEL; SENEFF (1986) propuseram os métodos de compensação: normalização espectral dependente das palavras chaves, do inglês *Codeword Dependent Cepstral Normalization*, subtração espectral, do inglês *Spectral Subtraction*, combinação de modelos paralelos, do inglês *Parallel Model Combination* e máxima verossimilhança para regressão linear, do inglês *Maximum Likelihood Linear Regression*, respectivamente. Esses métodos tinham como objetivo utilizar filtros para diminuir a interferência do sinal ruidoso antes de utilizar um reconhecedor automático da fala. Tais métodos assumem que os ruídos são estacionários, caracterizados por possuírem pouca variação ao longo do espectro do sinal. Em sistemas reais, geralmente, encontram-se ruídos não estacionários, ou seja, ruídos com grandes variações ao longo do espectro do sinal.

COOKE et al. (2001) e LIPPMANN (1997) utilizaram método de compensação capaz de lidar com ruídos não estacionários, chamado de características ausentes, do inglês *Missing-Feature*. Esse método foi inspirado no sistema auditivo humano, onde a redundância do sinal da fala pode ser utilizada para melhorar sua compreensão, ainda que grande parte do espectro do sinal esteja corrompido por ruído. Isso ocorre porque o componente auditivo, preferencialmente, processa os componentes de alta energia do sinal da fala, enquanto suprime a energia mais fraca do sinal (GONZALEZ et al., 2013).

O método de compensação de características ausentes é o principal grupo de métodos utilizados na reconstrução das características ausentes da fala. As regiões, ao longo do espectro do sinal, que contém uma grande quantidade de ruído e não foram possíveis identificar os componentes presentes nessas regiões, são denominadas de regiões ou componentes não confiáveis. As

demais regiões são chamadas de regiões ou componentes confiáveis e são utilizadas diretamente para reconstruir as características ausentes da fala (KIM; STERN, 2010).

Para encontrar essas regiões, diversos autores utilizam o oráculo. O oráculo é o perfeito conhecimento das amostras, sendo fornecido, previamente, a melhor probabilidade *a priori* da localização dos componentes confiáveis e não confiáveis do sinal. Para isso, é necessário ter o mesmo sinal com presença e ausência de ruído. Então, a diferença log-espectral entre o sinal limpo e o sinal ruidoso é utilizada para definir os componentes. A Equação 3.1 mostra como o oráculo é calculado.

$$O_m(t) = \begin{cases} 1 \text{ (Confiável), se } s_m(t) - \hat{n}_m(t) > \alpha \\ 0 \text{ (Não Confiável), caso contrário} \end{cases} \tag{3.1}$$

onde m é a frequência de banda no tempo t; α é o limiar definido experimentalmente, geralmente, tem o valor de 0,05; $s_m(t)$ é o sinal limpo; $\hat{n}_m(t)$ é a estimativa do sinal ruidoso, definido pela Equação 3.2.

$$\hat{n}_m(t) = \log(\exp(x_m(t)) - \exp(s_m(t))), \text{ se } x_m(t) > s_m(t); \tag{3.2}$$

onde $x_m(t)$ é o sinal ruidoso;

Neste capítulo, serão descritos vários trabalhos sobre reconstrução das características ausentes da fala, separando-os em duas partes. Na primeira, serão citadas as principais técnicas para identificar as regiões confiáveis e não confiáveis do sinal, utilizando o método de compensação de características ausentes. Na segunda, serão mostradas as técnicas para reconstruir as características ausentes da fala após a identificação das regiões.

3.1 Métodos de compensação de características ausentes

Para determinar as regiões confiáveis e não confiáveis do sinal, o método de compensação de características ausentes utiliza uma máscara. Segundo MOORE (1989), uma máscara é proveniente das características utilizadas, que auxiliam na identificação dos componentes mais intensos da fala, que dominam o sinal, auxiliando na distinção das regiões.

Para obter as máscaras do sinal da fala, BARKER et al. (2000) e ANDRES et al. (2011) propuseram o uso de uma estimativa do nível do ruído comparando com um determinado limiar. Essa estimativa foi feita através do uso da compensação sigmoidal, do inglês *sigmoid compensation*. O problema desse método é que o limiar foi definido utilizando apenas uma base e com intensidades e níveis de ruídos estabelecidos. Como o limiar não é ajustado dinamicamente, essa técnica falha quando aplicada em ambientes reais.

Basicamente, há duas abordagens para o método de compensação de características ausentes: a abordagem de marginalização (COOKE et al., 2001; BARKER; COOKE; ELLIS, 2005) e de imputação (SELTZER; RAJ; STERN, 2004; BADIEZADEGAN; ROSE, 2010). Na marginalização, são utilizados apenas os componentes confiáveis do sinal, descartando os componentes não confiáveis. Para isso, é feita uma integração dos componentes não confiáveis (Equação 3.3), utilizando uma distribuição marginal $p(x_r|C)$ em vez de $p(x|C)$, onde x representa os componentes do sinal, x_r componentes confiáveis e C a distribuição do sinal da voz:

$$p(x_r|C) = \int_{-\infty}^{\infty} p(x_u, x_r|C) dx_u \tag{3.3}$$

onde x_u são os componentes não confiáveis do sinal.

O principal problema da abordagem de marginalização é a utilização de apenas características espectrais (basicamente energia). Uma vez que, para reconstruir as características ausentes da fala, as características provenientes do *cepstrum*, por exemplo as extraídas pelo MFCC (discutida no Capítulo 2), apresentam um maior número de discriminantes que podem auxiliar quando as amostras contém ruído (RAJ; SELTZER; STERN, 2004). Outro problema dessas características está ligado ao Modelo Escondido de Markov (HMM). Para estimar a distribuição do modelo da fala do HMM, como por exemplo $p(x_r|C)$ da Equação 3.3, pode-se utilizar misturas de gaussianas completas ou covariância diagonal (WAHID et al., 2012), necessitando de muitos dados de treinamento para sistemas com grandes vocabulários.

A abordagem de imputação estima os componentes não confiáveis. Os métodos probabilísticos são os mais utilizados para essa estimação. SELTZER; RAJ; STERN (2004) e BADIEZADEGAN; ROSE (2010) propuseram o uso de classificadores probabilísticos para definir a máscara do sinal. Porém, segundo KIM; STERN (2011), a proposta feita pelos autores não apresenta boa taxa de reconhecimento da fala, quando o classificador Bayesiano é treinado com ruído branco e testado com vários outros tipos de ruídos. Para mitigar esse problema, KIM; STERN (2011) propuseram métodos de classificação dependente da frequência, do inglês *Frequency-Dependent Classification*, e ruído colorido, do inglês *Colored Noise*, justificando que a dificuldade dos autores que utilizavam a abordagem probabilística ocorreu porque a máscara definida fez com que os quadros de cada janela do sinal aumentasse o seu grau de influência em relação aos quadros mais próximos (região de vizinhança), diminuindo a generalização no reconhecimento das sentenças.

Nas Seções 3.1.1 e 3.1.2, serão discutidos os métodos de classificação dependente da frequência e ruído colorido. Esses métodos foram utilizados para comparação com esta pesquisa. A escolha se deu pela ampla comparação que os autores fizeram com trabalhos anteriores, nos quais obtiveram melhores resultados, e por definir uma máscara do sinal, utilizando características acústica, que servirá como base para os nossos experimentos.

Classificação Dependente da Frequência

KIM; STERN (2011) propuseram uma modificação no classificador Bayesiano com objetivo de isolar a influência dos quadros vizinhos, através da variação espectral dentro de cada quadro, simulando o efeito de vários tipos de ruído. Para isso, os autores fizeram uso de 12 coeficientes para regiões confiáveis e 11 coeficientes para regiões não confiáveis. Totalizando 23 coeficientes, que foram encontrados através da combinação dos coeficientes cepstrais, provenientes do descritor MFCC (discutido no Capítulo 2), com as técnicas medida de nivelamento espectral e filtros combinados.

Medida de Nivelamento Espectral

A medida de nivelamento espectral, do inglês *Spectral Flatness Measure* (SFM), verifica a tonalidade nas regiões do sinal, identificando os tons dominantes em cada quadro e mensurando a quantidade de ruído na amostra. Através das tonalidades, são identificadas as regiões com presença ou ausência da fala (JOHNSTON, 1988).

O cálculo do SFM é feito através da razão entre as médias geométricas e aritméticas do sinal, de acordo a Equação 3.4.

$$SFM(m) = \frac{\prod_n x_m(n)^{\frac{1}{N_m}}}{\frac{1}{N_m}\sum_n x_m(n)} \tag{3.4}$$

onde $x_m(n)$ é a magnitude espectral do sinal da fala de tamanho N_m.

Filtros Combinados

Devido à natureza dos harmônicos, a maior quantidade de energia em um sinal sem ruído, se concentra nos harmônicos. Quando há presença de ruído, essa energia é aumentada tanto nos harmônicos, quanto, principalmente, no restante do sinal. Assim, calculando a proporção entre a energia do sinal e a energia presente nos harmônicos, é possível estimar o nível de ruído presente na amostra. Para isso, KIM; STERN (2011) utilizaram o filtro de resposta ao impulso de duração infinita, do inglês *Infinite Impulse Response* (IIR), e o *pitch*, obtido através do algoritmo de detecção de *pitch*, baseado em histograma, do inglês *Histogram-Based Pitch Detection Algorithm* (SELTZER, 2000) (Seção 3.1.1.3). A função de transferência do filtro IIR pode ser vista na Equação 3.5:

$$Hcomb(z) = z^{-p}/(1 - gz^{-p}) \tag{3.5}$$

onde $p = 1/F_o$ é o *pitch* do sinal, F_o é a frequência, e g é o parâmetro responsável por determinar o limiar dos harmônicos, definido experimentalmente.

Para capturar as energias do sinal que estão localizadas entre os harmônicos, é necessário fazer uma inversão do ganho da função de transferência, $Hcombshift(z)$, da seguinte maneira:

$$Hcombshift(z) = -z^{-p}/(1+gz^{-p}) \tag{3.6}$$

Utilizando essas funções, os autores estimam o filtro combinado, do inglês *Comb Filter Ratio* (CFR), a partir da razão logarítmica entre $Hcomb(z)$ e $Hcombshift(z)$. Quanto maior for esse resultado, menor será o nível de ruído no sinal. A Equação 3.7 mostra o cálculo do CFR.

$$CFR[i,w] = 10\log_{10}(\frac{\sum_{n_i} Ycomb[n_i,w]^2}{\sum_{n_i} Ycombshift[n_i,w]^2}) \tag{3.7}$$

onde $Ycomb$ e $Ycombshift$ são as saídas do sinal obtidas a partir do índice da janela i, da quantidade de janela que o sinal foi dividido n_i e do tamanho da janela w após o uso do $Hcomb$ e $Hcombshift$, respectivamente.

Após o uso dos descritores, KIM; STERN (2011) distribuíram os 12 coeficientes para representar regiões confiáveis da seguinte forma: 5 coeficientes mel-cepstrais, 5 coeficientes delta, 1 coeficiente fornecido pela técnica medida de nivelamento espectral e 1 coeficiente proveniente do uso de filtros combinados. Os mesmos coeficientes das regiões confiáveis são usados para definir as regiões não confiáveis, com exceção do coeficiente CFR. Totalizando apenas 11 coeficientes.

Uma vez que as características foram obtidas, a técnica baseada em agrupamento para reconstrução de características ausentes (RAJ; SELTZER; STERN, 2004), do inglês *cluster-based missing-feature reconstruction*, foi utilizada para reconstruir as características ausentes da fala. Na Seção 3.2.1.1 será discutido o método que foi replicado nos nossos experimentos.

Algoritmo de Detecção de *Pitch* Baseado em Histograma

O *pitch* é uma característica acústica que fornece a percepção auditiva do tom (altura do som em relação a um ponto de referência). KIM; STERN (2011) utilizaram o algoritmo de detecção de *pitch* baseado em histograma, sugerido por SELTZER; RAJ; STERN (2004), que serve para identificar o *pitch* da fala. Além do trabalho de SELTZER; RAJ; STERN (2004), outros trabalhos como HESS (1983) e RABINER et al. (1976), também propuseram métodos para identificar o *pitch* do sinal da fala. As diferenças, entre as técnicas, estão no nível de ruído que as amostras foram submetidas.

Para detectar o *pitch* através do histograma, múltiplas estimativas foram utilizadas e, através da regra da maioria, foi escolhida a melhor estimativa para as regiões onde há presença da fala. A regra da maioria foi definida da seguinte maneira: se uma única frequência ocorrer 25%

ou mais na largura de banda, essa região será marcada como presença da fala, caso contrário, será marcada como região ausente de fala. A Figura 3.1 mostra as quatro etapas utilizadas para a construção da técnica.

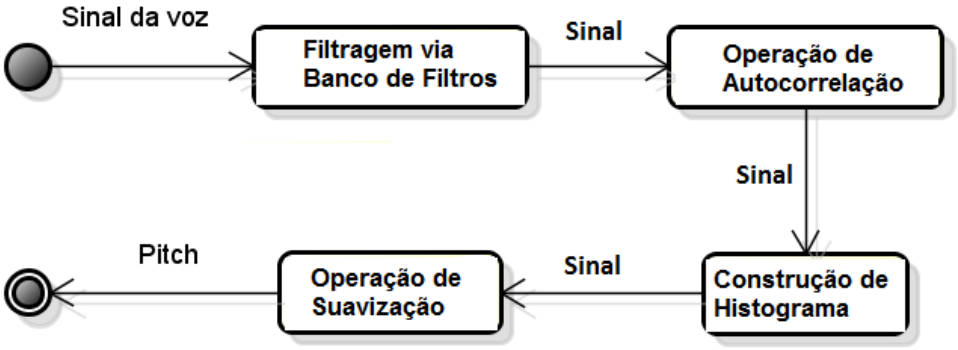


Figura 3.1: Etapas para a construção do algoritmo de detecção de *pitch* baseado em histograma.

A primeira etapa, filtragem via banco de filtros, foi baseada em SENEFF (1988). Nesta etapa utiliza-se um banco de filtros composto por 40 filtros combinados. O objetivo desta etapa é simular o sistema auditivo periférico que é responsável pela separação dos sinais (MEDDIS; HEWITT, 1991; WANG; HARPER, 2002). Com isso, acredita-se que a representação do espectro ficará próxima de um sinal sem ruído.

O modelo cascata, utilizado para o desenvolvimento do banco de filtros, pode ser visto na Figura 3.2. Esse modelo foi definido experimentalmente por SELTZER; RAJ; STERN (2004):

1. Para o filtro A, o filtro de resposta ao impulso de duração finita, do inglês *Finite Impulse Response* (FIR), é composto por 8 zeros.

2. Para os 40 filtros B, o filtro FIR é composto por 2 zeros.

3. Para os 40 filtros C, o filtro IIR é composto por 4 pólos e 4 zeros.

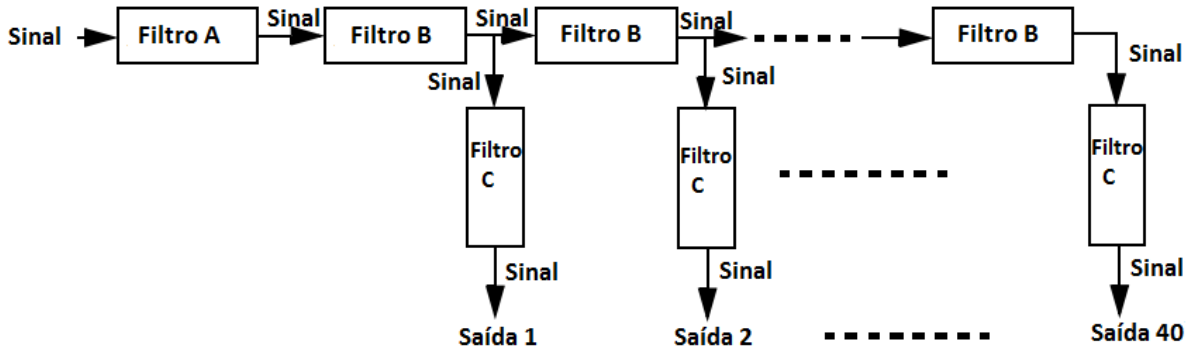


Figura 3.2: Cascata utilizada na implementação do banco de filtros. Adaptada de SELTZER (2000).

As Figuras 3.3 e 3.4 mostram a primeira e a última saída da cascata do banco de filtros, respectivamente, aplicado ao sinal da fala *"She had your dark suit in greasy wash water all year"*. Na Figura 3.3, as frequências mais altas sofreram poucas atenuações e as de baixas frequências, muita vezes marcadas como ruído, dominam parte do sinal. Já na Figura 3.4 as frequências mais altas foram atenuadas e as de baixas frequências tiveram o seu domínio no sinal suprimido, através do uso dos filtrados.

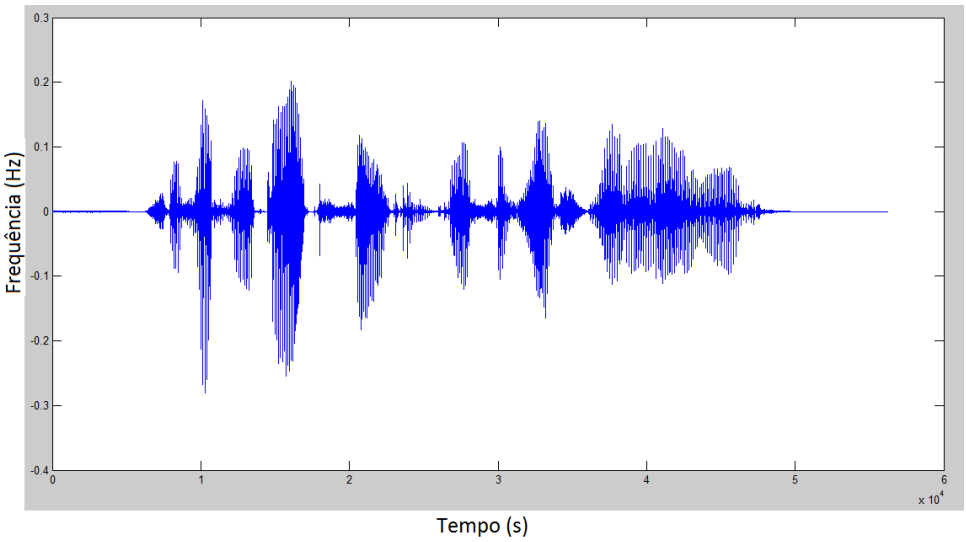


Figura 3.3: Primeira saída da cascata de filtros do sinal *"She had your dark suit in greasy wash water all year"*.

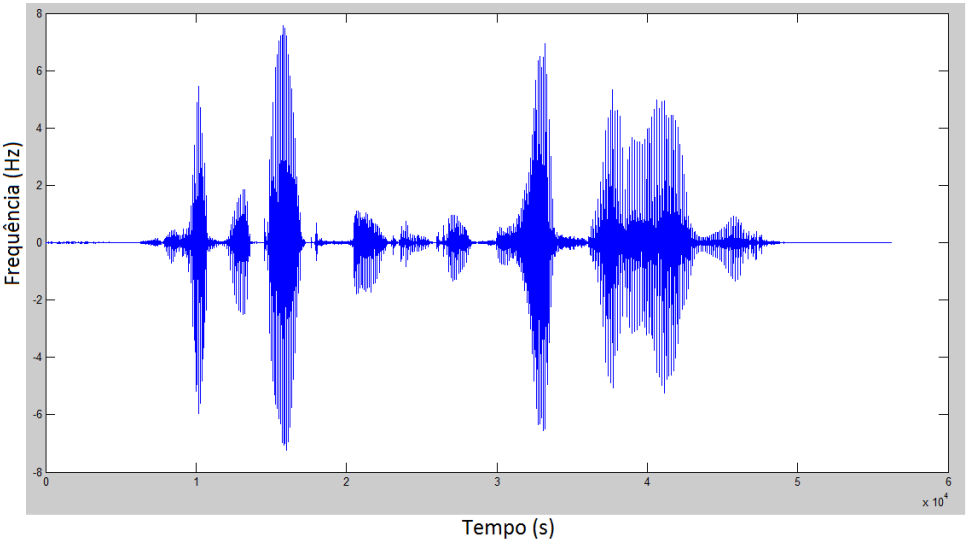


Figura 3.4: Última saída da cascata de filtros do sinal *"She had your dark suit in greasy wash water all year"*.

Após o uso do banco de filtros na amostra, os autores utilizaram 40 saídas representando o sinal. Cada saída representa uma janela no sinal, possuindo, assim, seu próprio harmônico.

Para eliminar picos falsos e para estimar o *pitch* em cada janela, a autocorrelação foi utilizada junto com o detector de envelope. A autocorrelação é a correlação cruzada de um

sinal com ele mesmo. Neste contexto foi utilizado um filtro que permite a passagem de baixas frequências, o filtro passa-baixa, na saída da autocorrelação. Um exemplo do comportamento do sinal, após essa etapa, pode ser visto através da Figura 3.5. Essa figura mostra a saída do sinal *"She had your dark suit in greasy wash water all year"* para a última janela.

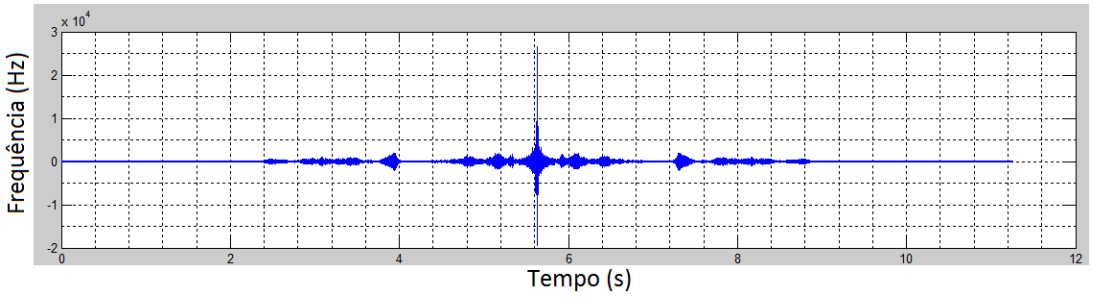


Figura 3.5: Autocorrelação da última janela.

As Figuras 3.6 e 3.7 mostram os envelopes do mesmo sinal, da primeira e última janela, respectivamente.

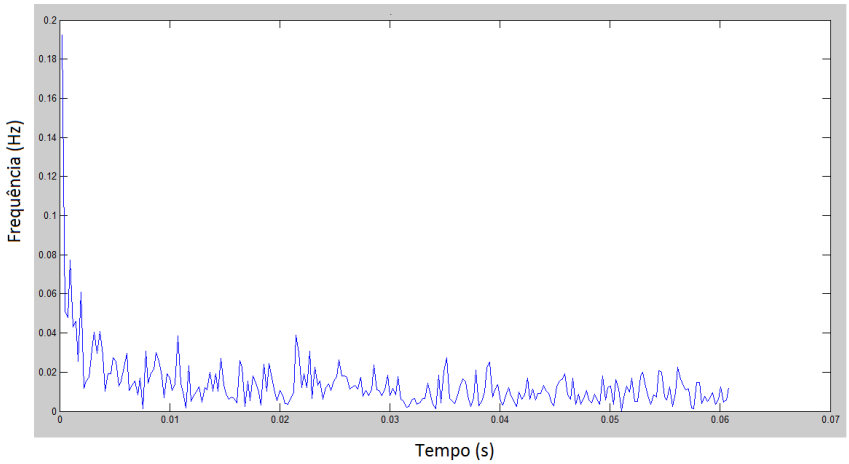


Figura 3.6: Envelope da primeira sub-banda.

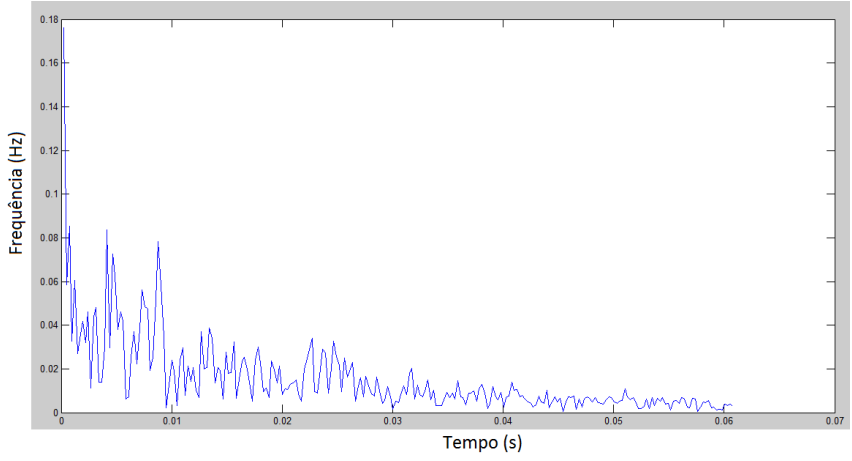


Figura 3.7: Envelope da última sub-banda.

O *pitch* em cada envelope é diferente. O mesmo é determinado através do cálculo das distâncias entre os dois maiores picos, resultando em uma estimativa do *pitch* para cada janela.

A terceira etapa é geração do histograma. Para isso, 40 *pitches* (um para cada janela) foram estimados na amostra, no intuito de identificar onde existem regiões com presença da fala e onde não há. Alguns desses *pitches*, principalmente aqueles com baixa frequência, podem representar a pausa entre palavras, nesse caso seria um *pitch* composto apenas por ruído, pois o algoritmo não utiliza segmentação da palavra ou fonema. Por essa razão, para as regiões sem presença da fala, a distribuição do sinal tende à uniformidade ao longo do tempo, enquanto que, para as regiões com presença da fala, não há uniformidade.

SELTZER; RAJ; STERN (2004) definiram um limiar para determinar se a região é vozeada ou não. Para tal, os autores utilizaram os seguintes critérios:

1. Região vozeada – Uma única frequência deve ocorrer em 25% ou mais na largura de banda.

2. Região não vozeada – Uma única frequência não ocorre em 25% ou mais na largura de banda.

Através desses critérios, foram definidos rótulos (presença da fala/ausência da fala) para as 40 estimativas de *pitch* da amostra. A Figura 3.8 mostra o histograma dos 40 candidatos a *pitch* da amostra *"She had your dark suit in greasy wash water all year"*.

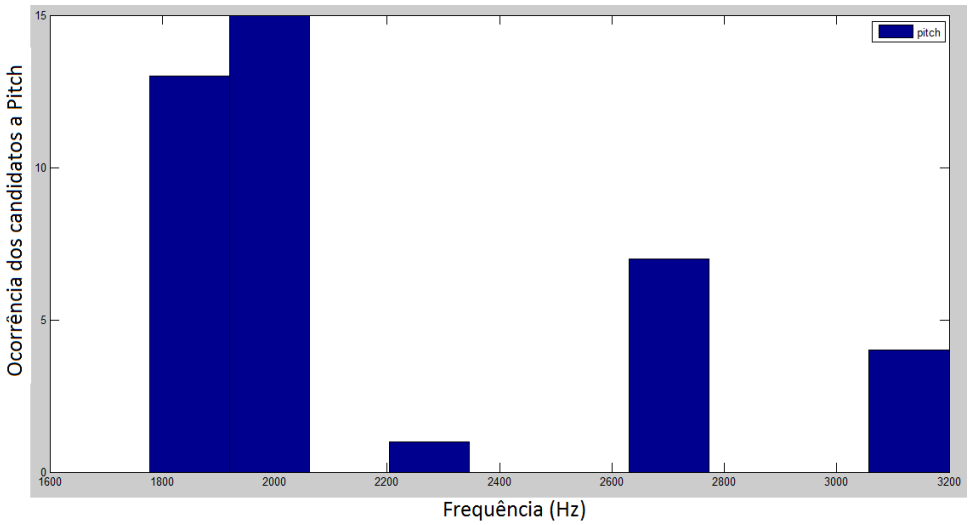


Figura 3.8: Histograma dos 40 candidatos a *pitch* da amostra.

Após a etapa do histograma é feita a última etapa, suavização no sinal. Este processo é utilizado para re-avaliar os rótulos definidos na etapa da construção do histograma. Essa reavaliação é necessária para evitar que um rótulo marcado como “região não vozeada”, esteja muito próximo da região rotulada como “região vozeada”. Isso ocorre, principalmente, com as últimas vogais da palavra, onde a energia do sinal é menor do que as vogais pronunciadas no meio da palavra. A suavização foi definida utilizando a seguinte regra:

1. Para ser considerada região com presença da fala, deve haver 3 janelas consecutivas marcadas como “região vozeada”. Caso não haja, será marcada como “região não vozeada”.

2. Para ser considerada “região não vozeada”, deve haver 3 janelas consecutivas marcadas como “região não vozeada”. Caso não haja, será marcada como “região vozeada” e o *pitch* será determinado pela interpolação linear com o seu vizinho mais próximo. Portanto, se as janelas marcadas como “região não vozeada” estiverem dentro do limiar de 8Hz (definido experimentalmente), essas janelas serão marcadas como “região vozeada”.

Após a realização dessas etapas, o *pitch* encontrado é passado para o cálculo da função de transferência Hcomb(z), Equação 3.5, e Hcombshift(z), Equação 3.6. Obtendo, assim, a máscara do sinal.

3.1.1 Ruído Colorido

O uso do ruído colorido foi proposto por KIM; STERN (2011) com o objetivo de diminuir a interferência dos quadros adjacentes. O objetivo era encontrar um maior grau de variabilidade temporal e espectral do sinal contaminado e, ao mesmo tempo, usando uma máscara capaz de funcionar independentemente dos tipos de ruídos.

Baseado nisso, cada banda da frequência Mel do sinal, conversão da escala Hertz para a escala Mel (Capítulo 2), foi dividida em N partições onde, cada partição, não interfere em nenhuma outra. Essa técnica foi chamada de método de partição multi-banda, do inglês *multi-band partition method*, para geração do ruído colorido (KIM; STERN, 2011).

No método de partição multi-banda, cada N partições da frequência Mel pode conter ruído, gerando até 2^N tipos de ruídos coloridos para cada banda do sinal. Para gerar cada partição, o ruído colorido não utiliza o filtro passa-faixa, como em KIM; STERN (2006), e sim manipula diretamente as frequência em cada banda. Para isso, o ruído branco é removido do sinal no domínio da frequência e então é utilizada a inversa da transformada de Fourier no restante da amostra, para obter as partições.

A combinação, para gerar o ruído colorido, é feita aleatoriamente, utilizando os mesmos intervalos de tempo para cada quadro. A Figura 3.9 ilustra o uso do ruído colorido na amostra *“eight seven two nine seven two six oh”* com ruído a 0dB contaminado pelo ruído do tipo carro, dividido em 23 bandas (janelamento) de frequência Mel, com quatro partições em cada frequência, combinada a cada 30ms.

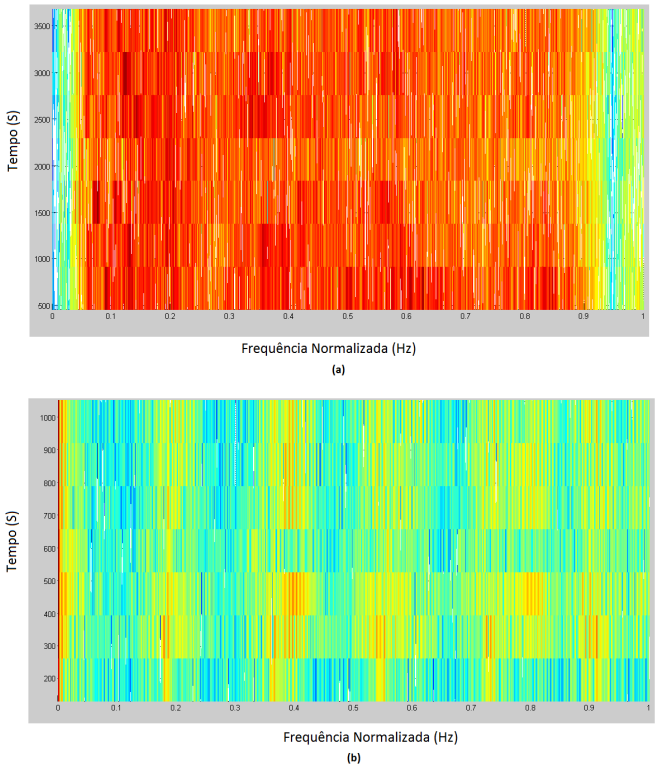


Figura 3.9: Espectrograma do sinal da voz apresentado nas Figuras (a) sinal na frequência Hertz com ruído do tipo carro a 0dB (b) resultado do sinal após o uso da técnica ruído colorido.

3.2 Métodos para Reconstrução da Fala

A máscara do sinal fornece as características que serão utilizadas na reconstrução da fala. Para reconstruir as características ausentes da fala, os métodos probabilísticos e as redes neurais, com treinamento supervisionado, são os mais utilizados. Recentemente, métodos com treinamento não-supervisionado foram utilizados para reconstruir as características ausentes da fala. LI; SIM (2014) utilizaram as técnicas de baixo ranking e decomposições esparsas no espectrograma ruidoso, comparando-as com técnicas não-supervisionadas, para estimar a amostra limpa, sem conhecimento prévio do sinal, e assim reduzir o ruído no sinal.

A diferença de um treinamento supervisionado para o não-supervisionado é que, no treinamento supervisionado, há um conhecimento do padrão de entrada e sua respectiva saída. Entretanto, no treinamento não-supervisionado existe somente o conhecimento do padrão de entrada. As abordagens supervisionadas são utilizadas para problemas de classificação e regressão. As abordagens não-supervisionadas são utilizadas para agrupar ou associar dados (HAYKIN, 2001).

3.2.1 *Métodos Probabilísticos*

O Modelo Mistura de Gaussianas, do inglês *Gaussian Mixture Model* (GMM) (HAYKIN, 2001), é uma das formas de modelagem de séries temporais mais comuns para reconhecimento da fala, com ou sem reconstrução de características ausentes (RABINER; SCHAFER, 2007). Essa série consiste na combinação linear de funções gaussianas de diferentes densidades de probabilidade. Esse modelo tem como vantagem a baixa exigência de espaço computacional (memória) e o baixo volume de dados para armazenar as informações, quando se trabalha com grandes conjuntos de dados.

Ao longo dos anos, a combinação do GMM com o HMM se tornou o modelo mais utilizado para o reconhecimento da fala (RABINER; SCHAFER, 2007). Isso ocorreu devido ao uso da máxima verossimilhança, do inglês *Maximum Likelihood Estimates* (MLE) (DEMPSTER; LAIRD; RUBIN, 1977), também conhecido como algoritmo de Baum-Welch [1], no treinamento do HMM, a fim de minimizar o risco da perda da informação e, consequentemente, para capturar a trajetória dinâmica da fala. A principal ideia para essa combinação é o fornecimento prévio da probabilidade *a priori* da transição e emissão do modelo. Vários autores propuseram essa combinação para reconstrução da fala utilizando métodos probabilísticos, dentre eles destacam-se:

1. RAJ; SELTZER; STERN (2004) utilizaram a máxima probabilidade *a posteriori* para determinar os componentes não confiáveis do sinal, através dos componentes confiáveis (abordagem descrita na Seção 3.2.1.1).

2. GONZALEZ et al. (2013) utilizaram o erro mínimo quadrado para explorar a correlação entre frequências. A proposta é semelhante a RAJ; SELTZER; STERN (2004), porém, com as expressões inseridas pelos autores, é possível observar a correlação entre as características de cada janela do sinal.

3. HARDING; MILNER (2011) utilizaram a máxima probabilidade *a posteriori* para estimar um envelope espectral que represente componentes confiáveis a partir de amostras ruidosas. Para tal, os autores utilizaram o descritor MFCC. A desvantagem desse método está na dificuldade em reconstruir características ausentes, quando o sinal contém ruídos não estacionários.

Na seção seguinte será abordado o método de agrupamento para reconstrução de características ausentes, proposto por RAJ; SELTZER; STERN (2004). Esse método foi utilizado para comparação neste trabalho.

[1] É um caso especial da maximização da expectativa, do inglês *Expectation–Maximization* (EM), utilizado para maximizar os parâmetros das funções. Esse algoritmo aprende as probabilidade de transição e as probabilidades de emissão do HMM, garantindo a convergência para um ponto de máximo local.

Agrupamento para Reconstrução de Características Ausentes

Após a definição da máscara do sinal da voz, RAJ; SELTZER; STERN (2004), propuseram um método estatístico baseado em agrupamento para reconstruir os componentes da fala.

Neste método é utilizada a máxima probabilidade *a posteriori* (BISHOP, 2006), para definir os componentes não-confiáveis do sinal, através dos valores dos componentes confiáveis do mesmo sinal. Para isso, o sinal sem a presença de ruído no domínio log-spectral, X, é dividido em k grupos. Assume-se que, para cada grupo, há uma distribuição gaussiana. A distribuição desse grupo é dada de acordo a Equação 3.8.

$$P(X|k) = \frac{\exp(\frac{-1}{2}(X-\mu_k)^T \theta_k^{-1}(X-\mu_k))}{\sqrt{(2\pi)^d|\theta_k|}} \tag{3.8}$$

onde X representa o sinal; d representa a dimensionalidade de X; μ_k e θ_k representam a média e a covariância dos k grupos, respectivamente.

A distribuição dos vetores espectrais é feita utilizando mistura de gaussianas de acordo a Equação 3.9.

$$P(X) = \sum_{k=1}^{K} c_k P(X|k) \tag{3.9}$$

onde c_k é a probabilidade *a priori* de k grupos.

Considera-se que o vetor do sinal ruidoso, Y, é subjacente ao vetor do sinal limpo, X. Portanto, o componente confiável, X_r, é semelhante ao componente observado em Y_r. Contudo, para os componentes não confiáveis, espera-se que o vetor que contenha ruído, Y_u, seja maior que o vetor correspondente X_u. Isso se dá pelo efeito aditivo que o ruído acrescenta nas amostras. Os k membros do grupo do sinal limpo são determinados pelas probabilidades *a posteriori* de acordo a Equação 3.10.

$$\hat{k} = argmax_k\{P(k)P(X|k)\} = argmax_k\{P(k)\int_{-\infty}^{Y_u} P(X|k)dX_u\} \tag{3.10}$$

Finalmente, os componentes não confiáveis, X_u, são reconstruídos utilizando as estimativas da máxima probabilidade *a posteriori* (MAP) com base nas regiões confiáveis, X_r, de acordo a Equação 3.11.

$$\hat{X_u} = argmax_{X_u}\{P(X_u|X_r, \mu_{X,\hat{k}}, \sum_{X,\hat{k}}, X_u \leq Y_u)\} \tag{3.11}$$

O problema desse método é a necessidade de ter o sinal limpo para cada locução. Em ambientes reais, nem sempre isso é possível.

3.2.2 *Redes Neurais Artificiais*

As redes neurais artificiais são sistemas paralelos compostos por unidades elementares, denominadas neurônios, que calculam determinadas funções matemáticas. Elas são capazes de aprender, memorizar e generalizar determinadas situações e problemas a elas apresentadas. Estas redes são classificadas de acordo a sua estrutura, conexões e treinamento (HAYKIN, 2001).

Segundo Lippmann (LIPPMANN, 1997), há duas razões para se utilizar redes neurais para o reconhecimento da fala. A primeira, é a capacidade de processamento paralelo ou distribuído, permitindo a utilização de grandes vocabulários em fala contínua. A segunda, são as capacidades de auto-adaptação, auto-organização e o auto-aprendizado de uma rede, que são capazes de criar modelos próprios para cada tipo de entrada, maximizando os resultados.

A combinação das redes neurais com os modelos probabilísticos proporcionou um rápido e eficiente treinamento. A dificuldade na combinação das técnicas está nas adaptações das redes neurais para fornecerem probabilidade *a priori* ou *a posteriori*. Para isso, modificações no arcabouço da rede, como por exemplo, o número de camadas escondidas, são necessárias para lidar com cada combinação (HERMANSKY; ELLIS; SHARMA, 2000).

HERMANSKY; ELLIS; SHARMA (2000) propuseram a combinação de duas abordagens utilizando a saída da rede MLP como entrada para o GMM, denominada de "Tandem". A Figura 3.10 mostra uma adaptação do fluxograma do "Tandem".

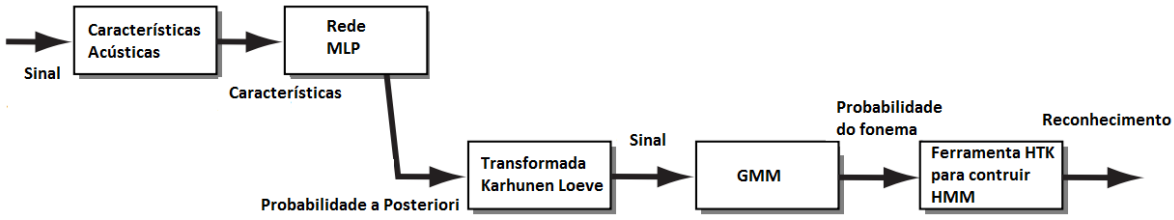


Figura 3.10: Fluxograma da combinação "Tandem". Adaptada de HERMANSKY; ELLIS; SHARMA (2000).

No fluxograma da Figura 3.10, a MLP fornece como saída a probabilidade *a posteriori* para cada fonema, que será descorrelacionada através do uso da transformada de Karhunen Loeve (KL) e enviada como entrada para o GMM. Através da ferramenta *Hidden Markov Model Toolkit* (HTK) [2] são construídos os estados da cadeia de Markov para reconhecer a fala. Como

[2]Disponível em: http://htk.eng.cam.ac.uk/, Visto em Março, 2016.

resultado, os autores mostraram que sua técnica apresentava 35% a mais de acerto, em relação ao uso do GMM de forma tradicional (sem realizar combinações).

Na última década, pesquisadores utilizaram redes neurais para reconstrução de características ausentes da fala. MASJOODI; VALI (2011), propuseram o uso da lógica fuzzy, combinada com o HMM e com o *Time Delay Neural Network* (TDNN) (CLOUSE et al., 1997) para reconstruir características ausentes da fala. Para isso, os autores propuseram uma modificação no método "agrupamento para reconstrução de características faltantes", onde, através da função sigmoide, é possível estimar a intensidade do ruído do sinal. No resultado, os autores obtiveram um taxa de acerto de 5% (utilizando o TDNN) e de 2% (utilizando o HMM) a mais comparado com o agrupamento para reconstrução de características faltantes. A Figura 3.11 mostra uma adaptação do fluxograma dessa proposta.

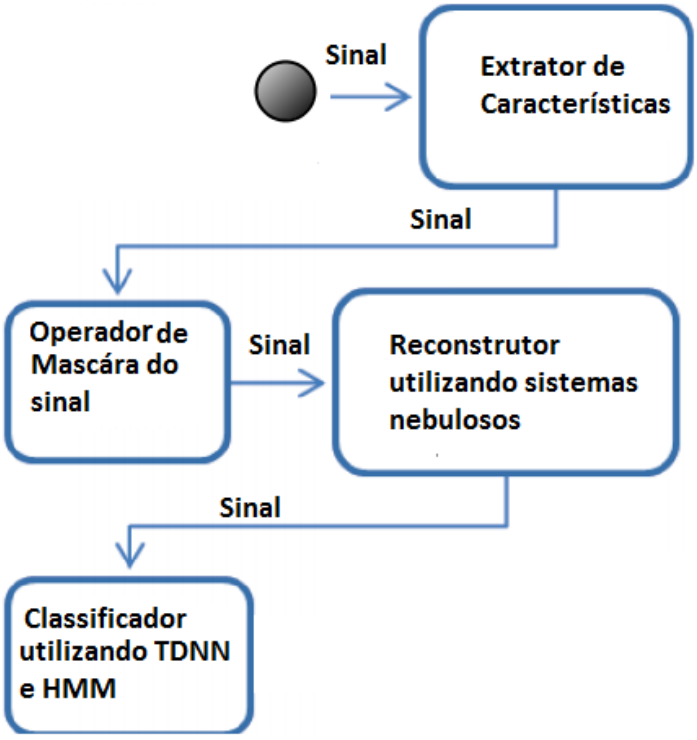


Figura 3.11: Fluxograma da técnica de reconstrução utilizando Fuzzy. Adaptada de MASJOODI; VALI (2011).

A primeira rede neural utilizando amplo vocabulário contínuo, surgiu em 2001 com o trabalho de SCHWENK (2007). Neste trabalho, o autor utilizou a rede *back-propagation* com duas camadas escondidas. A Figura 3.12 mostra uma adaptação da figura da rede utilizada por SCHWENK (2007). Nesta arquitetura, as entradas são normalizadas entre 0 e 1 e são utilizadas duas camadas escondidas. A primeira camada corresponde ao contexto de cada palavra, já a segunda, estima a probabilidade para cada palavra em uma abordagem supervisionada. Como resultado, o autor superou entre 0,4% a 1%, todas as técnicas que utilizaram como base o modelo transcrito pelo instituto nacional de padrões e tecnologia, do inglês *National Institute of Standards and Technology* (NIST). Essas técnicas tinham um erro médio de 14%.

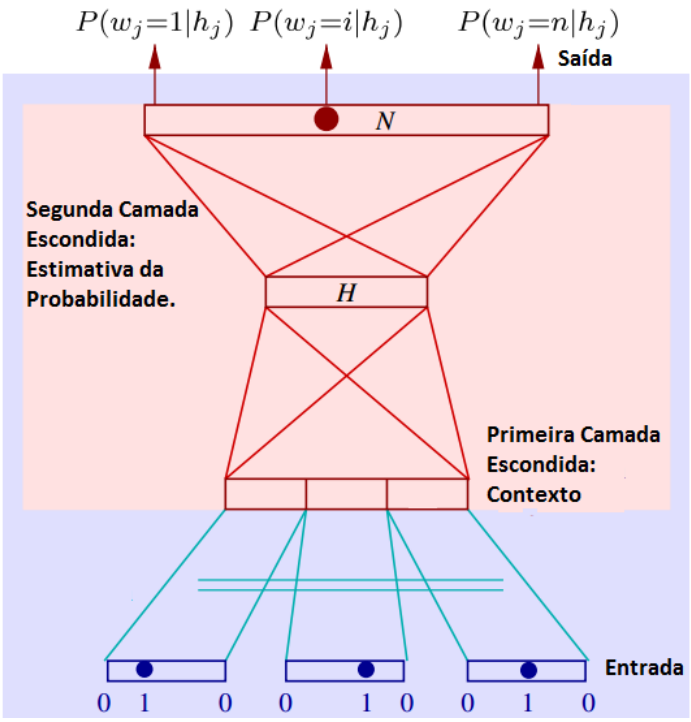


Figura 3.12: Arquitetura da primeira rede neural utilizando amplo vocabulário contínuo, onde $P(w_j = i|h_j)$ representa a probabilidade da palavra i dado o contexto h_j. Adaptada de SCHWENK (2007).

MIKOLOV et al. (2010) utilizaram redes neurais recorrentes aplicadas ao modelo de linguagem *n-gram*. As redes recorrentes têm como arquitetura a retroalimentação de informações que permite a criação de representações internas capazes de processar e armazenar informações temporais e sinais sequenciais. A Figura 3.13 mostra a adaptação da arquitetura da rede proposta por MIKOLOV et al. (2010). Com essa arquitetura, os autores obtiveram uma melhora de 18%, em relação às técnicas que utilizavam o modelo de linguagem de *n-gram* (cujo o erro médio era de 27%). Porém, essa proposta exigia grande poder computacional por ser um algoritmo custoso.

HORI; KUBO; NAKAMURA (2014) melhoraram o modelo proposto por MIKOLOV et al. (2010). Para isso, os autores pré-fixaram as possibilidades de cada palavra, não sendo necessário o uso de todo histórico da base para definir a probabilidade de cada palavra, diminuindo, desse modo, a carga computacional exigida e melhorando em 5% os resultados de MIKOLOV et al. (2010).

A partir de 2006, sugiram várias abordagens para reconstrução e reconhecimento da fala em tempo real utilizando aprendizagem profunda, do inglês *Deep Learning*. Uma das abordagens é a combinação da rede neural profunda, do inglês *Deep Neural Network* (DNN), com o modelo probabilístico GMM (LEI; LIN; HEIGOLD, 2013). Nesta abordagem os autores apresentaram um *framework* onde combinaram os erros do treinamento do GMM (associado ao HMM) com a rede neural profunda. Todavia, são utilizadas apenas características acústicas. Nos experimentos, os autores compararam os resultados da combinação proposta com o uso da rede DNN. Os

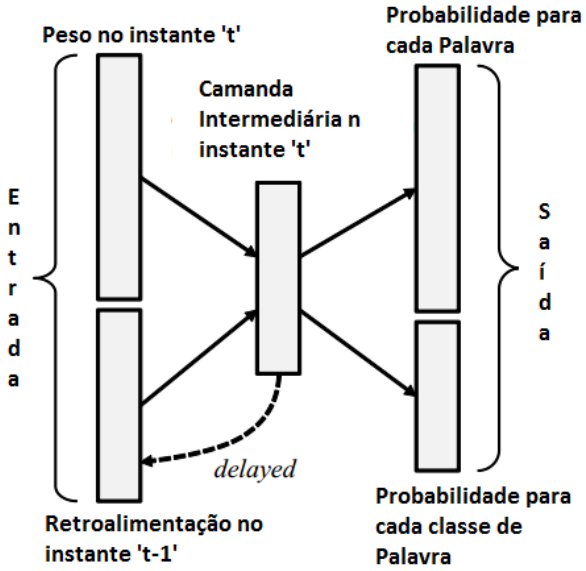


Figura 3.13: Arquitetura das redes neurais recorrentes aplicadas ao modelo de linguagem n-gram. Onde o instante t representa a palavra atual e o instante $t-1$ a última palavra apresentada ao modelo. Adaptada de MIKOLOV et al. (2010).

resultados mostraram que a combinação apresentou uma taxa de acerto superior ao uso da rede DNN, 5% a mais. A Figura 3.14 mostra a adaptação da arquitetura proposta por LEI; LIN; HEIGOLD (2013).

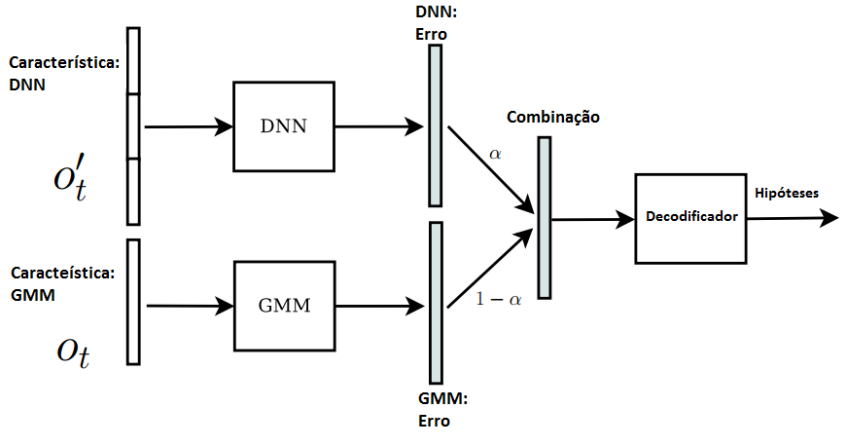


Figura 3.14: Arquitetura da combinação dos erros da rede neural profunda com o modelo de misturas gaussianas. Adaptada de LEI; LIN; HEIGOLD (2013).

GRAVES; MOHAMED; HINTON (2013) investigaram o uso das Redes Neurais Recorrentes Profundas, do inglês *Deep Recurrent Neural Networks* (DRNN), onde propôs a rede *Deep Long Short-Term Memory Recurrent Neural Network* que é uma combinação das arquiteturas das técnicas da Memória de Curto Prazo, do inglês *Long Short-Term Memory* (LSTM), com a rede neural recorrente, do inglês *Recurrent Neural Network* (RNN), para reconhecer fonemas na base TIMIT. Nos experimentos, os autores utilizaram apenas uma parte da base, chamada de

core test, e obtiveram uma taxa de erro médio de 17,7%.

LI; SIM (2014) propuseram o uso da Rede de Entrada Linear, do inglês *Linear Input Network* (LIN), da Máquina de Boltzman Restrita, do inglês *Restricted Boltzmann Machine* (RBM), e da rede DNN para reconhecimento da fala. Os autores compararam o desempenho da rede LIN para o reconhecimento da fala utilizando apenas características acústicas (MFCC), como também, o desempenho dessa rede para definir uma máscara do sinal combinada com as características acústicas para o reconhecimento da fala. Para isso, os sinais foram treinados utilizando RBM e os pesos foram passados como entrada para a primeira camada da rede DNN. Como resultados, os autores obtiveram uma taxa média de erro de 4,6%, utilizando a base Aurora 2, e 11,8%, utilizando a base Aurora 4.

Apesar dos visíveis avanços na área de reconhecimento da fala com reconstrução de características ausentes. Poucos trabalhos aplicados à reconstrução em tempo real foram encontrados. Um dos fatores prováveis é a difícil definição de uma máscara em um sinal corrompido com diversos tipos e intensidade de ruído e a definição do tipo de aprendizagem.

Um das possíveis soluções para resolver esse problema é combinar as caraterísticas acústicas e articulatórias, com um treinamento não-supervisionado, em um modelo que se adapte ao longo do tempo. Dessa forma, a rede se adapta ao ambiente, à medida que as amostras são apresentadas, exigindo menos recurso computacional do que a aprendizagem supervisionado quando aplicada ao amplo vocabulário contínuo. Isso acontece, pois não é necessário o armazenamento de todos os padrões previamente na memória. No Capítulo 4, será discutido o modelo proposto para reconhecer a fala com reconstrução de características ausentes.

4

Modelo Adaptativo para Reconhecimento da Fala

A reconstrução de características ausentes da fala em tempo real é uma área pouco explorada. Uma das causas possíveis, para este cenário, é a difícil escolha de qual modelo (características e categorizador) utilizar. Principalmente, quando os sinais são amostrados em diversos ambientes e níveis de ruídos distintos.

Neste trabalho, serão utilizadas: características acústicas, medida de nivelamento espectral e filtros combinados, para definir a máscara do sinal; características articulatórias, para auxiliar a rede neural no reconhecimento das caraterísticas ausentes do sinal; rede neural variante no tempo, LARFSOM (ARAUJO; COSTA, 2007), para o reconhecimento não-supervisionado.

O LARFSOM foi utilizado com o objetivo de explorar uma rede neural com aprendizagem incremental, não-supervisionada, na área de reconhecimento da fala com reconstrução de características ausentes. Este tipo de rede fornece a flexibilidade de adaptação ao ambiente, à medida que os padrões de entrada são apresentados, não precisando armazenar todos os padrões, previamente, na memória. Ou seja, a rede só cresce (em relação ao número de protótipos) se o padrão apresentado for um padrão que não se encaixa em um grupo já existente. Com isso, o modelo não demanda conhecimento prévio do número de grupos entre os padrões que serão fornecidos durante o treinamento (JAIN A.; MURTY, 1999). A Figura 4.1 mostra a arquitetura do modelo adaptativo proposto.

Nessa arquitetura, o sinal da fala é inicialmente pré-processado. Nesta etapa, o sinal é dividido em janelas, através do uso da técnica de janelamento de Hamming. Cada janela possui um tamanho de 256ms.

Esse sinal é então passado para os processamentos acústico e articulatório (em paralelo). No processamento acústico são extraídas as características provenientes do MFCC, medida de nivelamento espectral e filtro combinado. Essas características formam a máscara do sinal que será utilizada na identificação dos componentes confiáveis. No processamento articulatório são extraídas e identificadas as características da maneira de articulação do trato vocal. Para isso, foi utilizada a rede SOM para criar protótipos para cada característica articulatória, seis

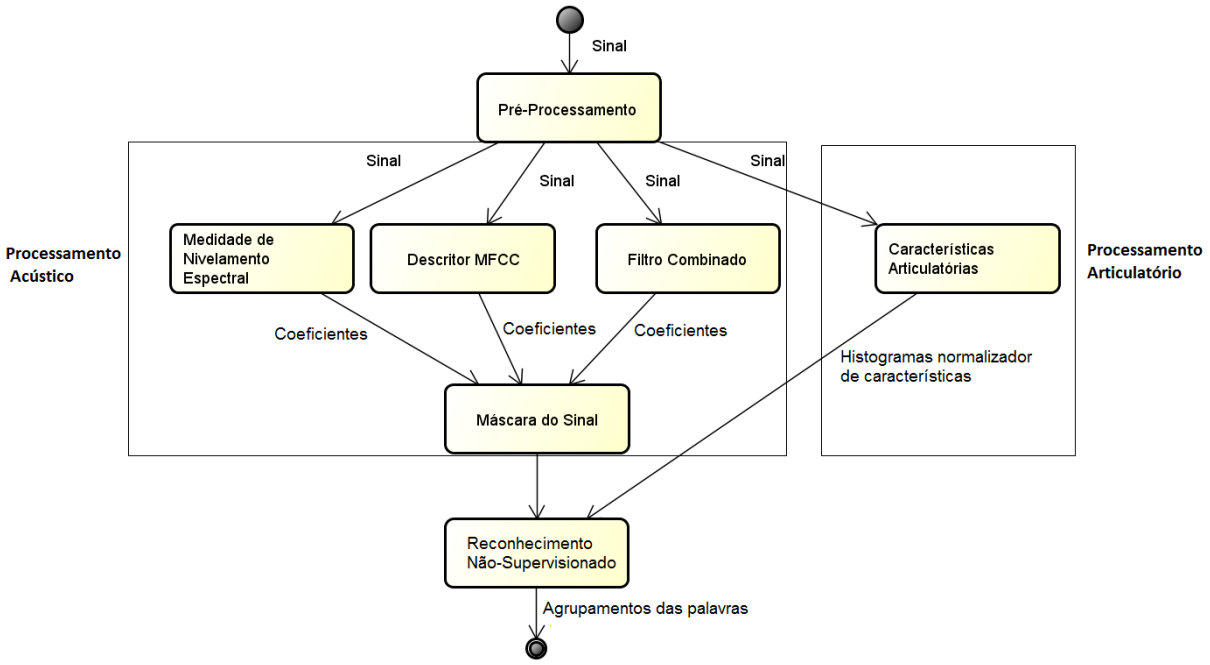


Figura 4.1: Arquitetura do Modelo Proposto.

neste trabalho. Como informação (saída) desta etapa, é construído o histograma normalizado das características que é passado para a etapa de reconhecimento não-supervisionado. Esse histograma fornece a quantidade de vezes que cada característica está presente no sinal.

A combinação da máscara do sinal e das características articulatórias formam um vetor de características que são passadas para a etapa de reconhecimento não-supervisionado. Nesta etapa, foi utilizada a rede LARFSOM. Essa rede reconstrói as características ausentes e reconhece a fala.

Esse tipo de modelo foi pouco explorado nos trabalhos anteriormente comentados. Tais modelos podem ser úteis para reconhecimento da fala em tempo real, pois, contrário ao modelo hierárquico, agrega os pontos em comum (comunalidade) sobre o relacionamento dos dados. Apresentando flexibilidade de adaptação e pouco armazenamento de memória.

Recentemente, as variações do Mapa Autos-Organizável (SOM) são as mais utilizadas no desenvolvimento de um modelo adaptativo. Uma rede SOM tem como estrutura uma camada de entrada e outra de saída. A camada de entrada recebe os valores de entrada e os retransmite para todos os nodos da camada de saída. Entretanto, na camada de saída, cada nodo está associado ao vetor de peso, interligados através das arestas, formando um mapa topológico. Para treinar essa rede, KOHONEN (1997) utilizou a aprendizagem não-supervisionada em um modelo competitivo, onde apenas o nodo que possui um vetor de pesos mais próximo do padrão da entrada fica ativo, e uma "função de vizinhança" que diminui ao longo das iterações, determinando nodos próximos ao vencedor e em que intensidade será o aprendizado. A Figura 4.2 mostra a estrutura de uma rede SOM adaptada de KOHONEN (1997).

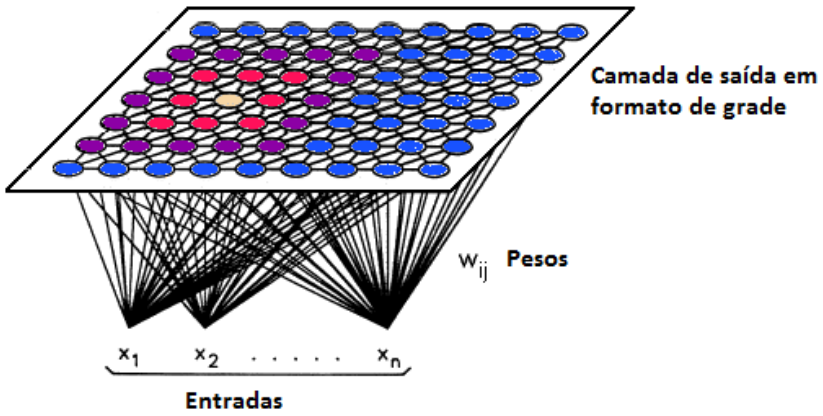


Figura 4.2: Estrutura de uma rede SOM. Adaptada de KOHONEN (1997).

Nas próximas seções, serão descritos os processamentos acústico e articulatório e como o LARFSOM foi utilizado para reconhecer a fala com reconstrução de características ausentes.

4.1 Processamento Acústico

As características acústicas (pitch, energia, estrutura dos formantes, número de picos e taxa de cruzamento por zero) foram utilizadas com o intuito de encontrar bons discriminantes que auxiliem na identificação das características ausentes da fala. Inicialmente, utilizou-se o descritor acústico MINERS para definir a máscara do sinal, entretanto, esse descritor apresentou dois problemas quando aplicado a diferentes bases:

- A definição dos limiares para os tons vermelhos (indicam componentes do sinal com presença de ruído) e verdes (indicam componentes do sinal com ausência de ruído) dependem da base. Não há uma equação que defina esses limiares, e, quando utilizado os mesmos valores em diferentes bases, o descritor apresentou uma taxa de erro de classificação de 52%.
- Quando existe identificação da amostra ruidosa, os filtros utilizados para minimizar o efeito do ruído acabam suprindo informações relevantes à localização das características ausentes.

Portanto, utilizou-se o processamento acústico semelhante a KIM; STERN (2011). Esta opção ocorreu devido aos resultados apresentados pelos autores quando o sinal contém ruído. Ao contrário do MINERS, o processamento foi definido, exclusivamente, para encontrar uma máscara com à finalidade de reconstruir as características ausentes. As características fornecidas por esse processamento são:

1. Descritor MFCC, com cinco coeficientes para cada amostra, resultantes da aplicação da Transformada Discreta de Cosseno (DCT);

2. Cinco coeficientes deltas;

3. Um coeficiente resultante da aplicação da técnica SFM;

4. Um coeficiente resultante da aplicação do filtro combinado.

Os coeficientes MFCC fornecem uma análise de características espectrais de tempo curto, baseando-se no uso do espectro da voz, convertido para uma escala de frequências mel. Esse descritor é o mais utilizado na área de reconhecimento da fala e, neste trabalho, foi utilizado para definir a máscara do sinal da fala. O fluxograma, para extrair os coeficientes MFCC, foi descrito no Capítulo 2, Figura 2.12.

O MFCC fornece coeficientes "estáticos" que capturam apenas mudanças temporais bruscas presentes no espectro. Para incorporar informações relativas às transições dos coeficientes estáticos entre quadros vizinhos, os coeficientes deltas foram utilizados. Esses coeficientes foram encontrados para cada amostra, de acordo a Equação 4.1.

$$d[t] = \frac{c[t+1] - c[t-1]}{2} \tag{4.1}$$

onde $d[t]$ são os coeficientes deltas computados no tempo t; c são os coeficientes do MFCC.

Por fim, foram utilizados os coeficientes SFM e filtro combinado. O SFM identifica as regiões com presença ou ausência da fala, enquanto que o filtro combinado define os níveis, SNR, de ruído nas amostras. Para isso, as Equações 3.4 e 3.7, Capítulo 3, foram utilizadas.

No total, são 12 coeficientes para regiões confiáveis e 11 para regiões não confiáveis, removendo o coeficiente proveniente da CFR, para definir a máscara do sinal. A limitação em utilizar apenas esse processamento é a não contemplação dos discriminantes articulatórios, pois as características articulatórias descrevem as propriedades da produção da fala e são robustas ao ruído.

4.2 Processamento Articulatório

Ao contrário das características acústicas, não há um descritor que forneça as características articulatória. Por isso, para identificar e extrair essas características, utiliza-se o mapeamento para cada fonema que é fornecida pelo IPA. Há três formas que podem ser utilizadas, junto com o mapeamento, para extrair as características articulatórias:

- Através de uso de equipamento para medir diretamente o trato vocal;
- Através da combinação das probabilidades de cada classe fornecidas pelas redes neurais;

- Através das características acústicas, utilizando a inversão de filtros, denominada de inversão da fala ou inversão acústica-articulatória.

Neste trabalho, escolheu-se a inversão acústica-articulatória para identificar e extrair as características articulatórias. Para isso, o pitch foi utilizado.

O pitch é uma característica acústica que fornece a percepção auditiva do tom (altura do som em relação a um ponto de referência). Ele foi utilizado para identificar as características relacionadas à maneira de articulação do trato vocal, através da função de diferença de magnitude média, do inglês *Average Magnitude Difference Function* (AMDF), proposto por TAN; KARNJANADECHA (2003).

O AMDF foi escolhido por apresentar uma baixa complexidade computacional, quando comparada com a função de autocorrelação, do inglês *Autocorrelation Function* (ACF) (HUANG; LEE, 2013), tanto quanto na função de correlação cruzada normalizada, do inglês *Normalized Cross Correlation Function* (NCCF) (JIAO; ZHIMI; CHAOCHENG, 2012). Isso ocorre porque, ao contrário das outras funções, o AMDF não utiliza multiplicação para extrair o pitch. A Função 4.2 mostra como o pitch é extraído utilizando AMDF. O pitch é o menor valor fornecido pelo AMDF.

$$Pk(t) = \frac{1}{N}\sum_{n=1}^{N-t} |x(t) - x(n+t)| \tag{4.2}$$

onde $Pk(t)$ é o pico da amostra no instante t; $x(t)$ é o sinal da fala; $x(n+t)$ são as amostras deslocadas no domínio do tempo no instante t; N é o tamanho da janela.

As características relacionadas à maneira de articulação, encontradas nos experimentos, foram: oclusiva, fricativa, aproximante, nasal, africada e vogal. O Algoritmo 1 descreve a extração do pitch do sinal da fala.

Algoritmo 1: Pitch

Inicialização: Crie subconjuntos (sd); Escolha o tamanho da janela de Hamming (N);
for $sd \leftarrow 0$ **to** sd_{max} **do**
    Dado o vetor de entrada ξ de sd, defina a janela de Hamming de acordo a Equação 2.2
    Encontre os picos do sinal, utilizando AMDF, de acordo a Equação 4.2
    Encontre o pitch P através do menor valor de AMDF:

$$P(t) \leftarrow min(Pk(t))$$

end

No Algoritmo 1, para cada característica articulatória (6 neste trabalho), foi criado um subconjunto (sd) contendo amostras dos sinais ruidosos e limpos. Assim, um especialista neural foi utilizado para aprender protótipos de características acústicas que compartilham

características articulatórias. Durante o reconhecimento, os especialista mais ativos determinarão as características articulatórias presentes.

A rede SOM foi utilizada para identificar as características articulatórias, especialista neural. Para isso, a rede criou protótipos para cada características e forneceu como informação, através da normalização do histograma, quais e quantas vezes cada característica ocorreu no sinal. A quantidade de vezes que cada característica está presente em um sinal auxilia na identificação das características ausentes do sinal, pois, se a mesma sentença é passada para rede, variando SNR ou locutor, o LARFSOM será capaz de reconstruir as características ausentes, baseando-se nos protótipos presentes na rede.

Para criar os protótipos articulatórios, as etapas de treinamento e teste são necessárias.

4.2.1 Treinamento

Nesta etapa, cada protótipo foi criado utilizando informação do *pitch* para cada característica. A Figura 4.3 mostra o fluxograma desta etapa.

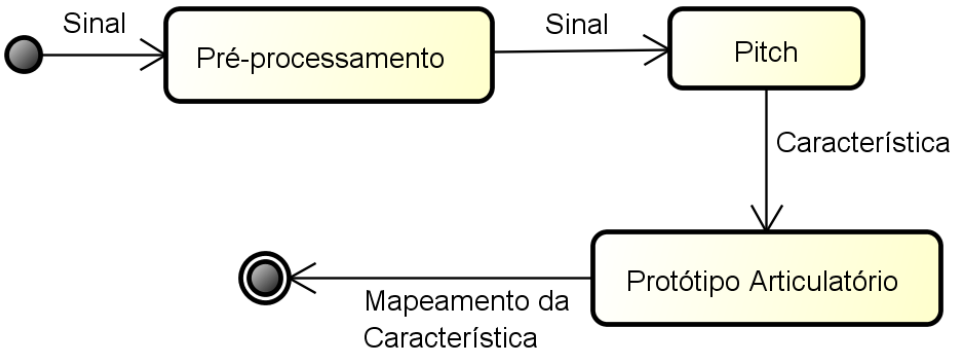


Figura 4.3: Fluxograma utilizado na Etapa de Treinamento.

De acordo a Figura 4.3, a etapa de treinamento é divida em três etapas:

1. Pré-processamento: Nesta etapa, o sinal é divido utilizando janela de Hamming. Cada janela possui um tamanho de 256ms.

2. Extração do *pitch* para cada característica: O *pitch* será utilizado para criar os protótipos articulatórios, extraídos de acordo ao Algoritmo 1.

3. Criação dos protótipos articulatórios: A rede SOM é utilizada para criar os protótipos. Como nas bases existem apenas seis características articulatórias, em relação a maneira de articulação, a rede SOM criou 6 protótipos, um para cada característica.

O Algoritmo 2 descreve como a rede SOM foi utilizada para criar os protótipos articulatórios. O algoritmo é semelhante a rede SOM tradicional. Os parâmetros utilizados nesse algoritmo estão descritos na Seção 5.2.

Algoritmo 2: Etapa de Treinamento: Protótipos Articulatórios

Inicialização: Inicialize, randomicamente, os valores iniciais dos pesos (w_j), escolha o número de épocas ($nepochs$), escolha o vetor de treinamento para cada subconjunto ($ntrain$), escolha a taxa de aprendizagem (ρ), escolha a taxa de decaimento exponencial da taxa de aprendizado (ρ_{decay}), vizinhança gaussiana(T), Variância inicial da função gaussiana ($sgm0$), Taxa de decaimento da variância gaussiana ($sgmdecay$), número de subconjuntos (sd_{max});
for $sd \leftarrow 1$ **to** sd_{max} **do**
    Escolha o subconjunto(sd);
    Amostragem: Apresente o vetor de treinamento (pitch) P para cada subconjunto sd;
    for $nepochs \leftarrow 1$ **to** $nepochs_{max}$ **do**
        Definir a taxa de aprendizagem: $\rho \leftarrow \rho * exp(-nepochs * \rho_{decay})$;
        Definir a variância da função gaussiana: $sgm \leftarrow sgm0 * exp(-nepochs * sgmdecay)$;
        for $ntrain \leftarrow 0$ **to** $ntrain_{max}$ **do**
            Definir a distância Euclidiana quadrática entre o vetor de treinamento (t) e cada neurônio do mapa: $d_j(t) = \sum_{i=1}^{i_{max}} (P_i - w_{ji})^2$;
            Encontrar o neurônio vencedor: $Win(t) \leftarrow min(d_j(t))$;
            Aplicar a equação de atualização de peso: $\Delta w_{ji} \leftarrow \rho(t) * T_{j,Win(t)}(t) * (P_i - w_{ji})$;
        end
    end
end

4.2.2 Teste

Nessa etapa, é feita a normalização do histograma para identificar o número de ocorrência de cada característica presente no sinal da fala. A Figura 4.4 mostra o fluxograma desta etapa.

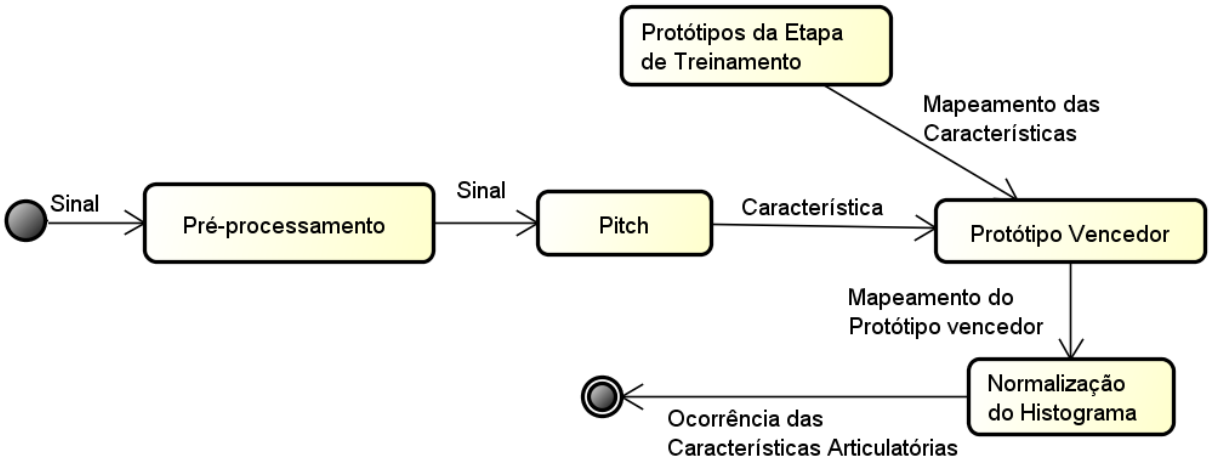


Figura 4.4: Fluxograma utilizado na Etapa de Teste.

De acordo a Figura 4.4, a primeira (pré-processamento) e a segunda (*pitch*) etapas são as mesmas da etapa do treinamento. Isso é necessário, pois se trata de um sinal que não foi processado na etapa de treinamento.

Na terceira etapa (protótipo vencedor) é calculada a distância euclidiana entre os protótipos provenientes da etapa de treinamento e o *pitch*. O protótipo vencedor é aquele com a menor

distância em relação ao padrão de entrada.

Na última etapa (normalização do histograma) é calculada as ocorrências de cada característica articulatória ao longo do sinal. Essas ocorrências auxiliam o LARFSOM na reconstrução das características ausentes.

O Algoritmo 3 descreve como a normalização do histograma foi criada. Os parâmetros utilizados nesse algoritmo estão descritos na Seção 5.2.

Algoritmo 3: Etapa de Teste: Normalização do Histograma

Inicialização: Inicialize, aleatoriamente, o padrão de entrada (X), inicialize o vetor de peso (w_{ij}) da etapa de treinamento, informe o número de sinais da base (dt), escolha o tamanho da janela de Hamming (N);
for $dt \leftarrow 1$ **to** dt_{max} **do**
    Dado um padrão de entrada (X) aplique a janela de Hamming com tamanho (N) ;
    for $N \leftarrow 1$ **to** N_{max} **do**
        Encontrar o pitch (P_{ij}) através da AMDF;
        Encontrar a distância euclidiana quadrática entre o pitch (P_{ij}) e cada neurônio do mapa (w_{ij});
        Encontrar o neurônio vencedor (Win_{ij}) ;
        Defina o valor 1 para a característica que representa o neurônio vencedor: $F(Win_{ij}) \leftarrow F(Win_{ij}) + 1$;
    end
    Defina o histograma: $H_{dt} \leftarrow F$;
end
Normalize o histograma: $norm(H)$;

4.3 Reconhecimento Não-Supervisionado

Nesta etapa, a máscara do sinal e as ocorrências de cada característica articulatória são passadas como entrada para a rede LARFSOM. Essa rede fará o reconhecimento não-supervisionado com reconstrução de características ausentes.

Diversas técnicas foram propostas baseadas na rede SOM. Dentre elas, destacam-se: as redes variantes no tempo gás neural crescente, do inglês *Growning Neural Gas* (GNG), e o crescimento da rede quando necessário, do inglês *Grow When Required* (GWR) (FRITZKE, 1995). O GNG gera uma estrutura multidimensional através das ligações, entre nodos, que refletem a topologia dos dados de entrada. Para isso, utiliza o aprendizado competitivo Hebbiano baseado nas redes GWR. Em resumo, as redes variantes no tempo têm como mecanismo:

- Habilidade de adicionar e remover nodos;
- Uma medida de erro local que é modificada a cada iteração e define o local de inserção/remoção de nodos;

- Conexões são inseridas entre os dois nodos com maiores respostas a um estímulo. Conexões demasiadamente velhas podem ser removidas, fator que induz ao aprendizado de características locais momentaneamente significativas;
- Permite a formação de agrupamentos em regiões distintas de outras, ou seja, sem conexões entre as mesmas.

O Mapa Auto-Organizável com Campo Receptivo Adaptativo Local (LARFSOM) (ARAUJO; COSTA, 2007) foi utilizado neste trabalho para reconhecer a fala e reconstruir as características ausentes. Essa técnica incorpora o aprendizado competitivo, a capacidade de agrupamento das redes SOM e o crescimento (quando necessário) das redes GWR, propondo uma estratégia para inserção de nodos, baseada na similaridade entre os padrões de entrada e os protótipos existentes na rede.

Segundo ARAUJO; COSTA (2009), as características da rede LARFSOM são:

1. A similaridade entre padrões de um mesmo grupo, que é determinada através de um limiar de ativação, que leva em consideração o campo receptivo da vizinhança local;
2. O campo receptivo é uma função de base radial (RBF), cujo objetivo é de limitar a influência dos agrupamentos entre os protótipos;
3. O crescimento da rede ocorre quando o valor do limiar de ativação do protótipo vencedor não for suficientemente alto em relação ao limiar de ativação;
4. Atualização dos pesos apenas do nodo vencedor;
5. A taxa de aprendizado é em função do número de vitórias do nodo, a qual se torna constante após um número máximo de vitórias;
6. Os nodos sem conexão são removidos após o treinamento.

O LARFSOM funciona para reconhecer e reconstruir as características ausentes, da seguinte forma:

1. Um padrão de entrada, constituído pela máscara e pelas ocorrências das características articulatórios no sinal, é recebido e, através da distância Euclidiana, é determinado o nodo vencedor, o nodo mais próximo do padrão de entrada;
2. Uma conexão entre as duas unidades mais próximas do padrão de entrada é criada;
3. O nodo vencedor é testado para saber se o seu protótipo é semelhante o suficiente com o padrão de entrada. Caso não seja, um novo nodo é inserido na rede (o peso desse nodo será o valor do padrão) e as duas menores arestas são criadas entre o novo nodo e as duas unidades mais próximas. Caso contrário, o nodo vencedor é ajustado com o intuito de aumentar a capacidade de responder a estímulos similares no futuro;

4. O processo é repetido até que um critério de convergência seja alcançado.

Essa rede tem como saída os agrupamentos para cada palavra. Através desses agrupamentos, o reconhecimento da fala é realizado. Os principais passos para criação do LARFSOM, segundo ARAUJO; COSTA (2009) são:

Passo 1: Inicialização de parâmetros: taxa de aprendizado final (ρ_f); modulador da taxa de aprendizado (ε); limiar de ativação (a_T) para pertencer a um grupo; número de vitória de cada nodo i $(d_i = 0)$; número máximo de vitórias de cada nodo (d_m); iteração (t); erro mínimo (e_{min}); e o número de unidades iniciais $(N = 2)$.

Passo 2: Apresentar o padrão de entrada (a escolha do padrão é aleatória) com peso igual ao valor do padrão.

Passo 3: Calcular a distância Euclidiana entre o vetor de entrada, ξ , e o vetor de pesos, w_i, de cada um dos nodos. A menor distância entre o vetor de entrada e todos os vetores de pesos de cada nodo, determina a unidade vencedora. Após isso, é incrementado o contador de vitórias da unidade vencedora $d_{S_1} = d_{S_1} + 1$.

Passo 4: Inserir uma conexão entre S_1 (nodo vencedor) e S_2 (segundo nodo mais próximo) caso não exista.

Passo 5: Calcular o campo receptivo de S_1: $r_{S_1} = \sqrt{(w_{S_1} - w_{S_2})^2}$

Passo 6: Calcular a atividade de S_1 por meio de uma Função de Base Radial: $a_{S_1} = \frac{exp(-||\xi - w_{S_1}||)}{r_{S_1}}$

Passo 7: Inserir um novo nodo se a ativação da unidade vencedora estiver abaixo do limiar (a_T), atualizando o número de nodos no mapa $N = N + 1$, removendo a conexão entre S_1 e S_2, calculando as distâncias do novo nodo em relação às duas unidades mais próximas: $d(w_{novo}, w_{S_1}), d(w_{novo}, w_{S_2}), d(w_{S_1}, w_{S_2})$; e inserindo conexão entre os nodos com as duas menores distâncias. Caso contrário, atualizar os pesos do nodo vencedor: $\Delta w_{S_1} = \rho \times (\xi - w_{S_1})$, onde ρ será calculado da seguinte forma:

$$\rho = \begin{cases} \varepsilon \times \rho_f^{d_i/d_m}, d_i <= d_m \\ \varepsilon \times \rho_f, d_i > d_m \end{cases}$$

Passo 8: Atualizar o número de iterações $t = t + 1$ e retornar ao passo 2, a não ser que o critério de parada: $e = \frac{1}{N} \sum_{i=0}^{N-1} ||w_i(t) - w_i(t+1)||^2 \leq e_{min} = 10^{-4}$ ou o número máximo de iterações tenha sido atingido.

Passo 9: Remover os nodos desconectados.

A taxa de aprendizagem do LARFSOM é definida experimentalmente para cada base. Entretanto, os experimentos foram conduzidos utilizando duas bases com os mesmos parâmetros, ou seja, sem ajustes manuais de parâmetros para cada base. Para isso, modificou-se o LARFSOM propondo uma nova equação que modifica, automaticamente, a taxa de aprendizagem:

$$\rho_f \leftarrow \frac{f}{\xi_n} \tag{4.3}$$

onde ρ_f é a taxa de aprendizagem, f é o número de característica e ξ_n é o número de padrões.

No LARFSOM, à medida que o nodos vai sendo inseridos no mapa, a busca pelos dois vencedores (etapa da competição) fica cada vez mais lenta (tempo computacional). Para aperfeiçoar essa busca, utilizou-se o algoritmo *Fast Nearest Neighbour Search* (FNNS) proposto por Cheung e Constantinides (CHEUNG; CONSTANTINIDES, 1993). Este algoritmo é baseado nas somas dos componentes do vetor de entrada e do vetor de peso. A ideia principal é dividir o vetor dos valores já adicionados dos nodos em duas partes: valores acima (soma dos componentes) ou abaixo do valor do vetor de entrada. O problema do FNNS é que o algoritmo encontra apenas um vencedor.

RêGO; FERREIRA; ARAúJO (2011) propuseram uma mudança no algoritmo de CHEUNG; CONSTANTINIDES (1993) para encontrar n vencedores. Essa proposta foi utilizada, neste trabalho, para encontrar os dois vencedores da rede LARFSOM, através dos seguintes passos:

Passo 1: Inicialização: Ordenar o vetor de peso de acordo a soma dos componentes: $A = S_0, S_1$, onde S_0 é o primeiro vencedor e S_1 é o segundo vencedor, encontrados através da soma das dimensões do vetor de peso.

Passo 2: Procura: Dado um vetor de entrada ξ, calcula-se a soma dos componentes desse vetor e cria-se o conjunto B da seguinte maneira: $B = \{S_{b_0}, S_{b_1}\}$,tal que: $||\sigma S_{b_0} - \sigma S_\xi|| < ||\sigma S_{b_1} - \sigma S_\xi||$, onde S_{b_0} e S_{b_1} são os índices dos nodos.

Passo 3: Encontrar a distância entre o vetor de entrada e o segundo vencedor (d_{min}) através da raiz quadrada da distância Euclidiana ($d(.)$), tal que: $d_{min} = d(w_{S_{b_1}}, \xi)$. Definir o limiar da pesquisa: $t = k \times d_{min}$, onde k é o número de dimensões. Iniciar os índices: $i_{up} = i_{down} = b_0$ e as bandeiras: $up_{search} = down_{search} = true$.

Passo 4: Caso $up_{search} = true$ e $i_{up} < A - 1$, faça:

1. $i_{up} = i_{up} + 1$
2. Caso $(\sigma S_{i_{up}} - \sigma_\xi)^2 \geq t$, faça : $up_{search} = false$
3. Caso contrário: $consider(S_{i_{up}})$

Passo 5: Caso $down_{search} = true$ e $i_{down} > 0$, faça:

1. $i_{down} = i_{down} - 1$
2. Caso $(\sigma S_{i_{down}} - \sigma_\xi)^2 \geq t$, faça : $down_{search} = false$
3. Caso contrário: $consider(S_{i_{down}})$

Passo 6: Função Consider: Essa função tem como objetivo atualizar o vetor B e, para o nosso caso, pode ser descrita de acordo ao algoritmo:

1. Função Consider(s)
   For ($i = 1$ até 2)
   If $d(w_{S_{b_i}}, \xi) < d_{min}$
   For ($j = 1$ até 2)
   $S_{b_j} = S_{b_{j-1}}$
   $S_{b_i} = s$

Passo 7: Repetir os passos 2 ao 5 até que up_{search} e $down_{search}$ receberem valores *"false"*.

Com o modelo definido, no Capítulo 5 serão discutidos os experimentos realizados através do modelo proposto. Tal modelo foi validado através do uso de duas bases que possuem diferentes propósitos.

5 Experimentos

Os experimentos realizados para validação deste trabalho foram divididos em três partes. Na primeira, replicaram-se os experimentos de KIM; STERN (2011), sem utilizar o oráculo, para verificar a influência da informação prévia da probabilidade *a priori* do sinal, em relação ao reconhecimento. Para isso, a base Aurora 2 (PEARCE; HIRSCH; EUROLAB, 2000) com e sem separação da relação sinal-ruído, foi empregada.

Na segunda parte, o modelo adaptativo proposto foi aplicado à base Aurora 2. KIM; STERN (2011), em seus experimentos, utilizaram apenas ruídos do tipo carro e balbucio da base Aurora 2. Com o intuito de comparar os resultados encontrados nesta tese com os descritos por KIM; STERN (2011), utilizamos, inicialmente, apenas esses dois tipos de ruídos. Em seguida, utilizamos todos os tipos de ruídos para comparar com os resultados presentes em LI; SIM (2014).

Na última parte, utilizaram-se a base TIMIT[1]. Para isso, os experimentos foram divididos em duas etapas: na primeira, toda a base foi utilizada. Na segunda, apenas os sinais disponíveis no conjunto "Core Test" foram utilizados.

5.1 Bases de Dados

As bases Aurora 2 e TIMIT foram utilizadas nos experimentos. A base Aurora 2 é constituída por números que foram gravados para a base TIdigit[2] e possui um conjunto de sinais limpos e ruidosos com uma taxa de amostragem de 8.000Hz. Ao todo, a base possui um conjunto de treinamento com 8.440 sinais sem ruído e três conjuntos de teste com ruído ("Teste A","Teste B" e "Teste C"), gravadas por 110 locutores, sendo 55 homens e 55 mulheres. Para cada nível de relação sinal-ruído (SNR) dos conjuntos de testes, existem 1001 sinais da fala. As relações sinais-ruído presente na base são: -5dB, 0dB, 5dB, 10dB, 15dB e 20dB gravadas dentro do metrô, carro, rua, restaurante, aeroporto, estação de trem, sala de exibição e balbucio.

A divisão dos conjuntos de treinamento ocorreu devido à forma de amostragem dos

[1]Disponível em: https://catalog.ldc.upenn.edu/LDC93S10, Visto em Março, 2017.
[2]Disponível em: https://catalog.ldc.upenn.edu/ldc93s1, Visto em Março, 2017.

ruídos. Tanto o Teste A quanto o Teste B tiveram os seus ruídos adicionados, seguindo as regras de transmissão definidas pela ITU-T (2001). Ambas foram gravadas com as seguintes SNRs: -5dB, 0dB, 5dB, 10dB, 15dB e 20dB. A diferença é que, no Teste A, os ruídos adicionados foram gravados nos seguintes ambientes: dentro do metrô, som de carro, sala de exibição e balbucio. As gravações do Teste B foram feitas nos ambientes: rua, restaurante, aeroporto e estação de trem. Em relação ao Teste C, tem duas diferenças, a primeira diz respeito à transmissão utilizada, desta vez não foi utilizada a G712 e sim o sistema de referência intermediária modificado, do inglês *Modified Intermediate Reference System* (MIRS) (GAO; SU, 2003). A segunda, está nos tipos de ruídos utilizados, foram gravados na rua e dentro do metrô.

A base TIMIT possui um total de 6.300 sinais da fala com uma taxa de amostragem de 16.000Hz. Cada sinal representa uma sentença, com tamanho variado, e foram gravadas por 630 locutores. Cada locutor pronunciando 10 sentenças no total de 5,4 horas de gravação. Essa base possui sotaques de 8 regiões dos Estados Unidos (Norte, Sul, Oeste, Norte de *Midland*, Sul de *Midland*, *New England* e *New York*). Ao todo 70% dos locutores são homens e 30% são mulheres (GAROFOLO et al., 1990).

A base é dividida em conjunto de treinamento e teste. Para isso, o balançamento fonético e a divisão equilibrada dos sotaques por conjunto foi observada. Porém, nenhuma sentença presente no conjunto de treinamento se repete no conjunto de teste. No total, o conjunto de treinamento possui 4.620 locuções pronunciadas por 462 locutores, possuindo, assim, no conjunto de teste 1.680 locuções pronunciadas por 168 locutores (GAROFOLO et al., 1990).

O uso dessas duas bases proporcionou a validação do modelo aplicado a locuções de sentenças de tipos, tamanhos, sotaques e locutores diferentes.

5.2 Parâmetros

O tamanho da janela de Hamming utilizada para o pré-processamento foi de 256ms. Esse tamanho foi definido experimentalmente.

Os parâmetros utilizados na rede SOM para criação dos protótipos articulatórios podem ser vistos na Tabela 5.1. Esses mesmos parâmetros foram utilizados nas bases Aurora 2 e TIMIT.

Tabela 5.1: Parâmetros utilizados na rede SOM

Parâmetro	**Valor**
Taxa de Aprendizagem (ρ)	0,02
Taxa de Decaimento da Taxa de Aprendizagem (ρ_{decay})	0,01
Variância da Função Gaussiana ($sgm0$)	10
Taxa de Decaimento da Variância da Função Gaussiana ($sgmdecay$)	0,02
Número de Épocas ($nepochs$)	30

Para identificar quais parâmetros da rede LARFSOM são mais sensíveis quando aplicado ao reconhecimento da fala com reconstrução de características ausentes. Empregou-se o método

de amostragem chamado “Hipercubo Latino”, do inglês *Latin Hypercube Sampling* (LHS) (KOLLIG; KELLER, 2002). O LHS é um método estatístico que gera valores de parâmetros, obedecendo a uma função de densidade de probabilidade dos dados. Com auxílio do LHS, verificou-se que o limiar da ativação é o parâmetro mais sensível da rede. Os parâmetros e seus respectivos valores utilizados na rede LARFSOM, após o uso do LHS, podem ser vistos na Tabela 5.2.

Tabela 5.2: Parâmetros utilizados na rede LARFSOM

Parâmetro	**Valor**
Limiar de Ativação (a_T)	7,64
Modulador da Taxa de Aprendizagem (ε)	0,3
Erro Mínimo (e_{min})	0,001
Número de Épocas ($nepochs$)	10

Com esses parâmetros, a rede apresentou poucas variações dos resultados (em relação à taxa de reconhecimento) quando treinada com diferentes bases, tipos e intensidades distintos de ruídos, necessitando de poucas épocas (apenas 10) para encontrar os melhores resultados.

5.3 Experimento 1: Classificador Dependente da Frequência

Os experimentos realizados por KIM; STERN (2011) utilizando o método classificador dependente da frequência, do inglês *Frequency-Dependent Classifier*, foram replicados. Em seus experimentos, os autores utilizaram o oráculo e apenas dois tipos de ruídos da base Aurora 2 (balbucio e carro), ruído branco da base “Noisex92” e ruído de música de uma base própria “10 músicas Coreanas”. A Tabela 5.3 apresenta a taxa de erro médio de classificação do discurso, encontrada pelos autores, utilizando a base Aurora 2.

Tabela 5.3: Erro Médio da Técnica Classificador Dependente da Frequência apresentado por KIM; STERN (2011)

	0dB	**5dB**	**10dB**	**15dB**	**20dB**
Carro	39,53%	16,76%	7,90%	5,13%	4,44%
Balbucio	46,70%	23,52%	10,55%	6,17%	3,81%

Com o intuito de verificar a influência do oráculo nos resultados dos autores, os experimentos desta tese foram feitos sem o emprego do oráculo. As Tabelas 5.4 e 5.5 mostram o erro médio e o melhor resultado, respectivamente, encontrados após 30 execuções do algoritmo e seus respectivos desvios-padrões.

Comparando os resultados da Tabela 5.3 com os das Tabelas 5.4 e 5.5, verifica-se que o oráculo tem influenciado no reconhecimento da fala. Isso acontece porque o oráculo fornece, previamente, a melhor probabilidade *a priori* do sinal limpo e ruidoso. Essa probabilidade é

Tabela 5.4: Erro Médio da Técnica Classificador Dependente da Frequência, sem utilizar o oráculo.

	0dB (σ)	**5dB (σ)**	**10dB (σ)**	**15dB (σ)**	**20dB (σ)**
Carro	75,00% (5,76)	60,00% (5,95)	57,00% (4,82)	48,00% (1,99)	30,00% (0,36)
Balbucio	79,00% (5,12)	74,00% (4,11)	69,00% (2,23)	65,00% (2,56)	55,00% (0,98)

Tabela 5.5: Melhor resultado (menor erro) da Técnica Classificador Dependente da Frequência sem utilizar o oráculo.

	0dB (σ)	**5dB (σ)**	**10dB (σ)**	**15dB (σ)**	**20dB (σ)**
Carro	70,00% (5,76)	58,10% (5,95)	54,04% (4,82)	46,90% (1,99)	29,10% (0,36)
Balbucio	74,62% (5,12)	70,30% (4,11)	67,90% (2,23)	63,10% (2,56)	54,00% (0,98)

utilizada para a reconstrução das características ausentes. Testes estatísticos não foram feitos devido os autores não apresentarem os desvios-padrões dos seus experimentos.

Com o objetivo de explorar a robustez ao ruído da técnica classificador dependente da frequência, sem o emprego do oráculo, foram utilizadas todas as intensidades (SNR) do ruído da base Aurora 2, sem separação, no treinamento e teste (exceto a intensidade -5dB que os autores originais não utilizaram). A Tabela 5.6 mostra o erro médio e o melhor resultado de classificação.

Tabela 5.6: Erro Médio e Melhor Resultado do Classificador Dependente da Frequência com intensidade de ruídos misturados, sem utilizar o oráculo.

	Erro Médio (σ)	**Melhor Resultado**
Carro	50,90% (4,88)	48,9%
Balbucio	53,50% (3,82)	50,87%

Na Tabela 5.6 foi observada que a técnica é sensível à variação da relação sinal-ruído. Esta dificuldade está relacionada às características utilizadas. São as características responsáveis por definir a máscara do sinal, que são extraídas através do descritor acústico MFCC (Capítulo 2). Esse descritor é largamente utilizado na área de reconhecimento da fala, porém, vários autores mostraram a ineficiência no reconhecimento da fala, quando existe a presença de ruído nos sinais, mesmo utilizando a combinação do GMM com o HMM, na reconstrução de características ausentes da fala.

5.4 Experimento 2: Modelo Proposto

O reconhecimento da fala com reconstrução de características ausentes foi feito utilizando o modelo adaptativo proposto nesta tese. Inicialmente, foram feitos experimentos utilizando apenas a máscara acústica do sinal, denominado de modelo acústico. Em seguida, foram feitos experimentos utilizando apenas as ocorrências das características articulatórias, denominado de modelo articulatório. Por fim, todo o modelo foi utilizado, ou seja, foram incorporadas as ocorrências das características articulatórias e a máscara do sinal nos experimentos, denominado

de modelo adaptativo.

5.4.1 Modelo Proposto: Avaliado na base Aurora 2

Nos primeiros experimentos com a base Aurora 2, apenas o modelo acústico foi utilizado. O objetivo é comparar esse modelo com os modelos probabilísticos.

As Tabelas 5.7 e 5.8 mostram a média de erro e o melhor resultado, com separação da intensidade de ruído, respectivamente. A Tabela 5.9 mostra os resultados sem separação da intensidade de ruído. Os números de grupos identificados e os números de nodos estão presentes nas Tabelas 5.7, 5.8 e 5.9. Os grupos e os nodos mostram como ficaram os mapeamentos dos sinais após o uso do modelo. Os números ideais seriam 1001 grupos e 1001 nodos inseridos, pois para cada intensidade de ruído, a base Aurora 2 possui 1001 sinais distintos, por isso, era esperado que cada sinal distinto fosse representando por um único nodo e que cada nodo representasse um grupo.

Tabela 5.7: Resultados do Modelo Acústico, com separação da intensidade de ruído, para os tipos de ruídos de carro e balbucio.

	-5dB (σ)	**0dB** (σ)	**5dB** (σ)	**10dB** (σ)	**15dB** (σ)	**20dB** (σ)
Carro	20,02% (3,43)	19,01% (3,66)	17,07% (3,21)	15,07% (2,63)	11,09% (2,58)	10,00% (2,40)
Grupos	830	832	832	840	853	859
Nodos	1330	1350	1352	1200	1220	1240
Balbucio	20,03% (2,10)	19,03% (2,65)	19,12% (2,72)	18,10% (3,34)	13,20% (3,01)	12,05% (3,17)
Grupos	822	826	826	828	846	850
Nodos	1300	1309	1389	1326	1360	1300

Tabela 5.8: Melhores Resultados do Modelo Acústico, com separação da intensidade de ruído, para os tipos de ruídos de carro e balbucio.

	-5dB	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**
Carro	18,00%	17,00%	14,00%	13,00%	10,00%	8,00%
Grupos	840	845	855	862	880	900
Nodos	1333	1320	1332	1232	1281	1299
Balbucio	18,00%	17,00%	15,00%	14,00%	11,00%	8,00%
Grupos	840	845	849	868	878	900
Nodos	1380	1329	1389	1306	1302	1296

A Tabela 5.7 mostra a média de erro e o melhor resultado sem separação da intensidade de ruído para o treinamento e teste, empregando os tipos de ruídos de carro e balbucio. Esses resultados foram publicados no trabalho de VIANA; ARAUJO (2016).

A rede SOM foi utilizada para identificar as características articulatórias através dos protótipos articulatórios. Para isso, a rede foi treinada com sinais limpos e ruidosos, contendo todos os tipos e intensidade de ruído presente na base.

Para avaliar a qualidade dos protótipos, foram utilizadas as métricas estatística: "precisão", do inglês *precion*, *recall* e a "medida *f*", do inglês *f-measure*. Essas métricas levam em

Tabela 5.9: Resultados do Modelo Acústico, sem separação da intensidade de ruído, para os tipos de ruídos de carro e balbucio.

	Erro médio (σ)	**Melhor Resultado**
Carro	20,00% (2,43)	15%
Grupos	830	850
Nodos	1330	1338
Balbucio	20,00% (1,90)	15%
Grupos	830	848
Nodos	1300	1324

consideração os seguintes aspectos:

- Verdadeiro-Positivo (VP): Ocorre quando a característica é agrupada corretamente em seu protótipo;
- Verdadeiro-Negativo (VN): Ocorre quando a caraterística não é agrupada em seu protótipo;
- Falso-Positivo (FP): Ocorre quando a característica é agrupada erroneamente em um determinado protótipo;
- Falso-Negativo (FN): Ocorre quando a característica não deveria ser agrupada em certo protótipo e a mesma, corretamente, não agrupa.

A precisão, definida pela Equação 5.1, leva em consideração o número de falsos-positivos encontrados durante os testes.

$$P = \frac{VP}{VP + FP} \tag{5.1}$$

onde P é a precisão, VP é o número de verdadeiros-positivos e FP o número de falsos-positivos.

O *recall*, definida pela Equação 5.2, leva em consideração o número de falsos-negativos encontrado durante o teste

$$R = \frac{VP}{VP + FN} \tag{5.2}$$

onde R é o *recall*, VP é o número de verdadeiros-positivos e FN o número de falsos-negativos.

A medida f, definida pela Equação 5.3, é a média harmônica entre a precisão e o *recall*.

$$F = 1 + B^2 \frac{P * R}{B^2 * P + R} \tag{5.3}$$

onde B foi definido como 1, P é a precisão e R o *Recall*.

A Figura 5.1 mostra a curva em relação a taxa de aprendizagem (ρ) da precisão, *recall* e medida f.

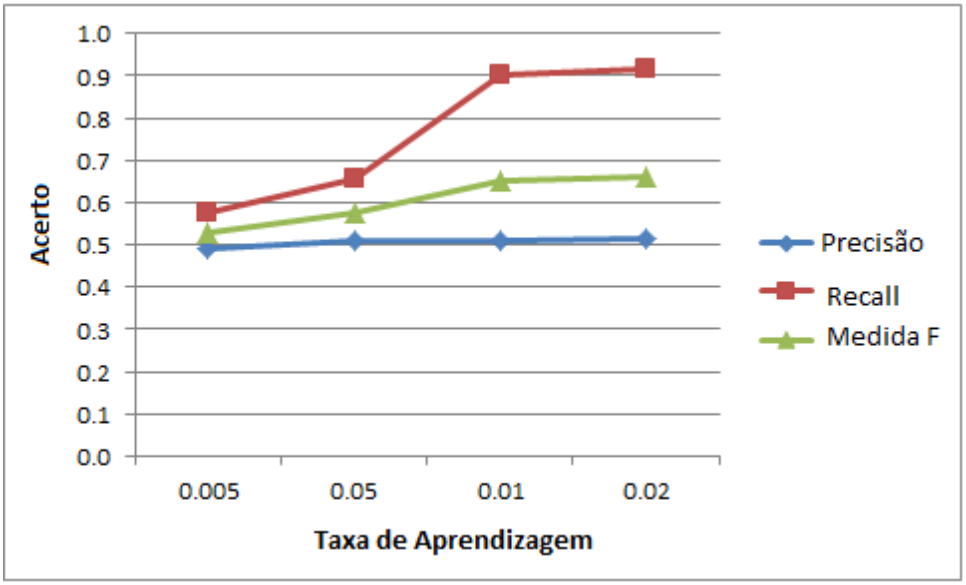


Figura 5.1: Medidas de Qualidade da rede SOM para a base Aurora 2.

Na Figura 5.1, a rede SOM apresentou poucos falsos-negativos, porém, em relação a precisão, o mesmo apresentou um elevado número de falsos-positivos. Isso interfere diretamente na reconstrução das características ausentes, pois, ao agrupar erroneamente uma características, o LARFSOM, responsável pela reconstrução, terá um nodo que não representa a caraterística correta.

Para verificar a importância da precisão sobre o modelo, apenas o modelo articulatório foi utilizado como característica. Nesses experimentos, toda a base Aurora 2 foi utilizada. Como resultados, foram obtidos uma taxa de erro médio e o melhor resultado, para o reconhecimento da fala, de 48% e 46%, respectivamente.

Utilizando o modelo adaptativo, os experimentos foram divididos em duas partes: na primeira foram utilizados apenas os ruídos de carro e balbucio com intuito de comparar os resultados obtidos com as técnicas *colored noise* (KIM; STERN, 2011), classificador dependente da frequência (KIM; STERN, 2011) e com o modelo acústico, proposto por VIANA; ARAUJO (2016). Na segunda parte, comparou-se o modelo proposto com as técnicas LIN-DNN (LI; SIM, 2014) e LIN-RBM-DNN (LI; SIM, 2014).

A Tabela 5.10 mostra o erro médio, o melhor resultado, o número de grupos e o número de nodos encontrados nos experimentos com o modelo adaptativo. Já as Figuras 5.2 e 5.3 mostram as comparações dos resultados do modelo adaptativo com as técnicas citadas anteriormente.

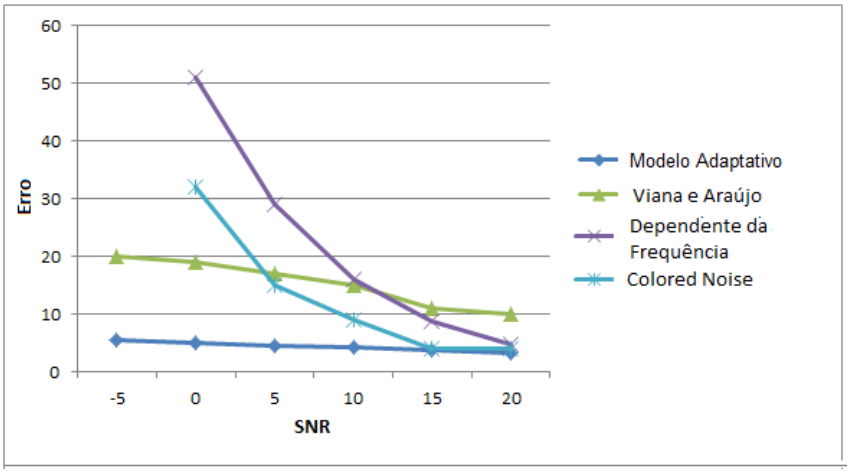


Figura 5.2: Média de Erro de Reconhecimento para o ruído do tipo Carro com SNRs: -5, 0, 5, 10, 15 e 20dB.

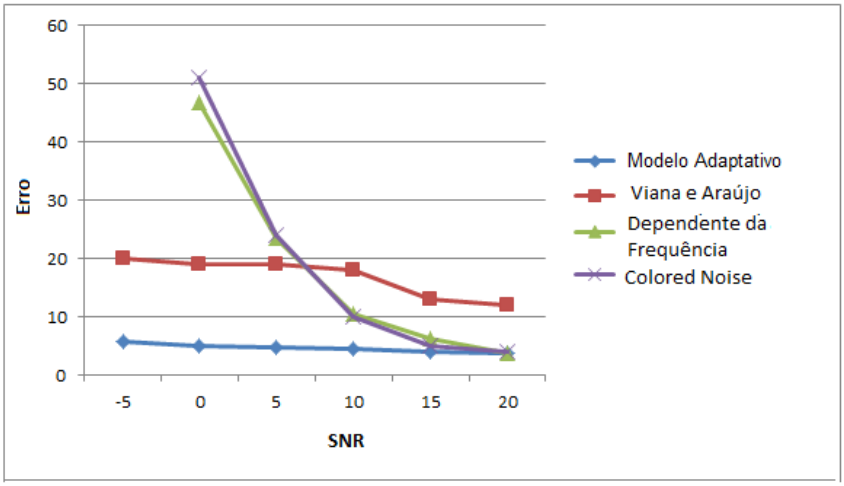


Figura 5.3: Média de Erro de Reconhecimento para o ruído do tipo Balbucio com SNRs: -5, 0, 5, 10, 15 e 20dB.

Para verificar a diferença (teste bilateral) e a relevância (teste unilateral) estatística dos resultados foi utilizado o teste de Wilcoxon, *signed-rank* (GIBBONS; CHAKRABORTI, 2011). Assumiu um intervalo de confiança de 95% e um nível de confiança (α) de 5%. Esse teste foi realizado apenas entre o modelo adaptativo (MA) e o modelo de VIANA; ARAUJO (2016). Pois, os outros autores, não descreveram os desvios-padrões de seus trabalhos. A Tabela 5.11 mostra o teste de hipótese utilizado no teste bilateral e unilateral.

A Tabela 5.11 fornece evidência estatística para sugerir que os métodos são diferentes e o modelo adaptativo apresenta menor erro de reconhecimento do que o modelo de VIANA; ARAUJO (2016).

Com o objetivo de verificar as diferenças entre sentenças gravadas por homens e mulheres no reconhecimento da fala, experimentos com separação de gênero foram realizados. A Tabela 5.12 mostra o erro médio, a média do número de grupos e a média do número de nodos para ruídos do tipo carro e balbucio separados por gênero.

Tabela 5.10: Erro Médio, Melhor Resultado, Número de Grupos e Número de Nodos para ruídos do tipo Carro e Balbucio.

Carro	**Média (σ)**	**Melhor Resultado**	**Grupos**	**Nodos**
-5dB	5,65% (1,54)	4,13%	970	1250
0dB	5,12% (1,43)	4,01%	984	1251
5dB	4,65% (1,65)	3,31%	986	1210
10dB	4,23% (1,32)	3,02%	986	1202
15dB	3,78% (1,12)	2,79%	988	1198
20dB	3,21% (1,11)	2,22%	993	1182
Balbucio	**Média (σ)**	**Melhor Resultado**	**Grupos**	**Nodos**
-5dB	5,76% (1,59)	4,32%	940	1256
0dB	5,14% (1,23)	4,41%	949	1251
5dB	4,89% (1,41)	3,39%	952	1243
10dB	4,54% (1,92)	3,12%	965	1201
15dB	3,92% (1,52)	2,81%	978	1187
20dB	3,76% (1,06)	2,65%	983	1187

Tabela 5.11: Teste de Wilcoxon, *Signed-Rank*, utilizando a base Aurora 2.

Teste Bilateral			
H_0	**H_1**	**P-Value**	**Conclusão**
MA = Viana e Araújo	MA $\neq$ Viana e Araújo	$1.62E^{-11}$	Rejeita H_0
Teste Unilateral			
H_0	**H_1**	**P-Value**	**Conclusão**
MA $<$ Viana e Araújo	MA $\geq$ Viana e Araújo	1	Não Rejeita H_0

Tabela 5.12: Erro Médio, Média do Número de Grupos e Média do Número de Nodos para ruídos do tipo Carro e Balbucio Separados por Gênero.

Carro	**Homem (σ)**	**Mulher (σ)**	**Grupos**	**Nodos**
-5dB	5,32% (1,79)	5,30% (1,61)	972	1250
0dB	4,89% (1,34)	4,80% (1,10)	984	1257
5dB	4,39% (1,75)	4,33% (1,45)	982	1211
10dB	4,19% (1,76)	4,08% (1,23)	983	1202
15dB	3,78% (1,12)	3,43% (1,11)	988	1198
20dB	3,21% (1,11)	3,01% (1,20)	993	1182
Balbucio	**Homem (σ)**	**Mulher (σ)**	**Grupos**	**Nodos**
-5dB	5,58% (1,65)	5,42% (1,21)	942	1216
0dB	5,01% (1,11)	4,98% (1,19)	939	1271
5dB	4,45% (1,23)	4,39% (1,42)	951	1243
10dB	4,54% (1,52)	4,12% (1,27)	962	1201
15dB	3,89% (1,23)	3,81% (1,77)	971	1206
20dB	3,72% (1,16)	3,65% (1,21)	991	1207

Os resultados apresentados na Tabela 5.12 em relação aos apresentados na Tabela 5.10 revelam uma acuracidade maior em alguns casos. Isso acontece porque o conjunto de treinamento

separado por gênero é menor do que o utilizado na Tabela 5.10, portanto há menos casos de variação, levando a tais acuracidades. Isso ilustra a robustez do modelo, em relação a taxa de erro, às variações por genêro e ao número de amostras, ocasionado pela natureza incremental do algoritmo.

Na segunda parte, foram empregados todos os sinais presentes nos subconjuntos Teste A, Teste B e Teste C da base Aurora 2. A Tabela 5.13 mostra o erro médio do modelo proposto, ao compará-lo com as técnicas LIN-DNN (LI; SIM, 2014) e LIN-RBM-DNN (LI; SIM, 2014).

Tabela 5.13: Erro Médio utilizando todos os sinais dos subconjuntos Teste A, Teste B e Teste C.

	Modelo Adaptativo (σ)	**LIN-DNN**	**LIN-RBM-DNN**
Teste A	4,51% (1,82)	4,30%	4,10%
Teste B	4,67% (1,76)	6,00%	5,70%
Teste C	4,21% (1,45)	5,30%	5,00%
Média	4,46% (1,67)	5,20%	4,93%

5.4.2 Modelo Proposto: Avaliado na base TIMIT

A base TIMIT foi utilizada para validar o modelo adaptativo em bases que utilizam falas contínuas. Essa base é mais utilizada no reconhecimento de fonemas, pois fornece a transcrição ortográfica e a localização, em relação ao tempo, de todos os fonemas presentes nos sinais. Neste trabalho a base TIMIT foi utilizada para o reconhecimento da fala.

Para avaliar o modelo adaptativo, dividiu-se os experimentos em duas partes: na primeira, todos os sinais dos conjuntos de treinamento e teste, foram colocados em um único conjunto denominado de "Base-Completa". Isso ocorreu porque nos conjuntos de treinamento e teste não existem sinais que se repetem, a divisão foi feita de acordo ao balanceamento fonético e não a partir de sentenças. Por isso, foi criado um único conjunto que foi posteriormente dividido em conjuntos de treinamento e teste com sinais que se repetem em ambos, isto é, ocorrências de uma mesma frase com diferentes locuções. Na segunda parte foram utilizados somente os sinais disponíveis no conjunto "*Core Test*". Para isso, foram criados subconjuntos de treinamento ("Treinamento A") e teste ("Teste A"). O conjunto *Core Test* possui 192 sentenças gravados por 24 locutores.

Da mesma maneira que foram realizados os experimentos com a base Aurora 2. Empregou-se a rede SOM para criar os protótipos para cada característica (com os mesmos parâmetros que a base anterior). Para avaliar a qualidade dos protótipos, as métricas estatística de precisão, *recall* e medida f foram utilizadas. Nas Figuras 5.4 e 5.5, as precisões, *recall* e as medidas f são mostradas, em relação a taxa de aprendizagem (ρ), para o conjunto *Core Test* e Base-Completa, respectivamente.

Assim como nos experimentos utilizando a base Aurora 2, a precisão apresentou um elevado número de falsos-positivos. Isso correu, devido ao uso dos mesmos parâmetros aplicados

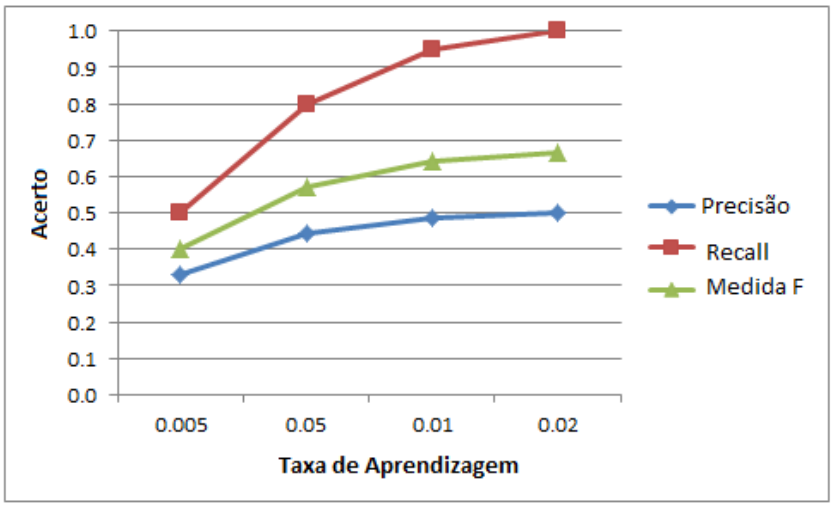


Figura 5.4: Medidas de Qualidade para a rede SOM utilizando o conjunto *Core Test*.

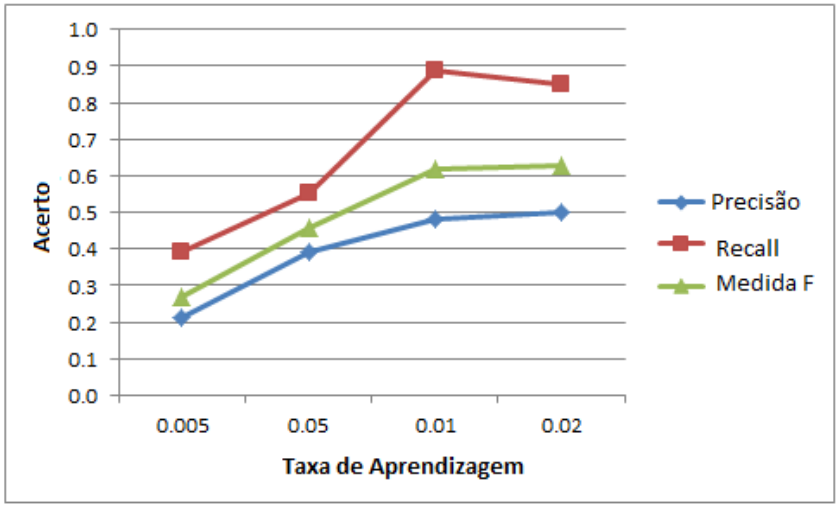


Figura 5.5: Medidas de Qualidade para a rede SOM utilizando a Base-Completa.

a duas bases distintas.

A Tabela 5.14 mostra o erro médio e o desvio padrão após executar o algoritmo 30 vezes. Não há nenhuma comparação com trabalhos anteriores, pois não foram achados trabalhos anteriores de reconhecimento da fala (e não de fonemas) com reconstrução de caraterísticas ausentes aplicado a essa base.

Tabela 5.14: Erro Médio para os conjuntos *Core Test* e Base-Completa.

	Modelo Adaptativo (σ)
Core Test	5,34% (1,05)
Base-Completa	8,36% (0,89)

5.5 Análise dos Resultados

Inicialmente, foi replicado o método classificador dependente da frequência, sem utilizar o oráculo. Verificando, desse modo, a influência de se conhecer a probabilidade *a priori* do sinal. Os resultados descritos na Tabela 5.6 ratificaram a sensibilidade do método em relação à variação da relação sinal-ruído, evidenciando a necessidade de se definir características robustas aos ruídos.

Com intuito de prover um novo modelo para o reconhecimento da fala, foi proposto o uso do modelo acústico, sem utilizar o oráculo. O problema desse modelo era a lentidão da rede LARFSOM. Essa rede apresentava lentidão, em relação ao tempo computacional, na etapa de competição. Nesta etapa, cada vez que os nodos são inseridos no mapa, mais lenta a técnica ficava. Para resolver esse problema, foi proposto o uso da técnica FNNS para encontrar os dois nodos vencedores. Com isso, foi reduzido em 80% o tempo de treinamento da rede.

As Tabelas 5.7 e 5.8 comparada com a Tabela 5.3, mostram que o modelo acústico, se comportou melhor, quando ruídos com intensidade de 0dB, foram apresentados. Para intensidade de 5dB, em média, os resultados para o tipo carro, apresentados por KIM; STERN (2011), foram próximos dos resultados descritos na Tabela 5.7. Já para o tipo de ruído de Balbucio, em média, os resultados obtidos foram melhores. Entretanto, esse modelo não foi melhor quando a intensidade de ruído vai diminuindo (10dB, 15dB e 20dB). Isso ocorreu porque não foi utilizado o oráculo, foi proposto o uso da aprendizagem não-supervisionada, não havendo, portanto, ajuste no limiar de ativação para cada intensidade de ruído. O objetivo desses experimentos eram encontrar parâmetros globais para o uso em sistemas em tempo real. KIM; STERN (2011) não fizeram experimentos para ruídos a -5dB.

Com o objetivo de verificar o comportamento do modelo acústico sem separação de ruído, realizou-se treinamento e teste com os ruídos de carro e balbucio juntos (Tabela 5.9). O modelo apresentou estabilidade (pouca variação nos resultados) à intensidade e tipo de ruído, sendo superior aos experimentos descritos na Tabela 5.6. Isso ocorreu porque o modelo foi capaz de reconhecer o padrão de entrada, mesmo contendo diferentes tipos de ruídos. Esse reconhecimento está ligado ao parâmetro a_T da rede LARFSOM, onde são verificas as diferenças entre o padrão de entrada e os protótipos já presentes na rede. Só é criado um novo grupo se a ativação da unidade vencedora estiver abaixo do a_T.

O modelo acústico apresentou uma baixa identificação do número de grupos, em média 838, e um elevado número de nodos, em média 1350. O ideal seria ter 1001 grupos e 1001 nodos. Essa baixa identificação correu devido a utilização dos mesmos parâmetros para bases distintas. Isso ocasionou erros de reconhecimento, pois algumas palavras deixaram de ser reconhecidas, devido a não presença do agrupamento.

No segundo momento, utilizou-se o modelo adaptativo, comparando com as técnicas *colored noise*, classificador dependente da frequência e com o modelo proposto por VIANA; ARAUJO (2016). A Figura 5.1 mostra a qualidade da rede SOM que foi empregada para

identificar as características articulatórias. A rede apresentou um número elevado de falsos-positivos, porém, um número baixo de falsos-negativos. Isso correu devido ao não ajuste dos parâmetros entre as bases, levando o modelo a agrupar algumas características, erroneamente, em um determinado protótipo.

A Tabela 5.10 mostra os resultados do modelo adaptativo aplicado a base Aurora 2. Esses resultados são melhores, em relação a taxa de reconhecimento, do que todos os experimentos anteriores, independente da intensidade do ruído. Além disso, o modelo apresentou pouca variação, em relação a taxa de reconhecimento, as intensidades e tipos de ruídos presentes na base. Isso aconteceu devido a robustez das características articulatórias em relação ao ruído.

O teste de Wilcoxon, *signed-rank*, foi empregado apenas em relação a Tabela 5.7, pois, os outros autores, não descreveram os desvios-padrões dos experimentos. O teste forneceu evidências estatísticas para sugerir que os métodos, Viana e Araújo (VIANA; ARAUJO, 2016) (Tabela 5.7) e o modelo adaptativo, são diferentes, como também, que o modelo adaptativo apresentou menor erro de reconhecimento do que o modelo proposto por VIANA; ARAUJO (2016).

A Tabela 5.13 compara o modelo adaptativo com as técnicas LIN-DNN e LIN-RBM-DNN. Como os autores não fornecerem os desvios-padrões, levou-se em consideração apenas as médias. No conjunto Teste A, a técnica LIN-RBM-DNN apresentou o menor erro médio de reconhecimento da fala. Porém, em relação aos conjuntos Teste B, Teste C e média geral, o modelo adaptativo apresentou melhores taxas de reconhecimento da fala. A vantagem do modelo, em relação as técnicas propostas por LI; SIM (2014), está nos poucos parâmetros necessários para ajustar a rede LARFSOM em relação a rede DNN.

A base TIMIT também foi utilizada para validar o modelo em bases que utilizam falas contínuas. As Figuras 5.4 e 5.5 mostram a qualidade da rede SOM, que foi empregada para identificar as características articulatórias, assim como, na Figura 5.1, a rede apresentou um número elevado de falsos-positivos, porém, um número baixo de falsos-negativos.

Nas pesquisas realizadas em trabalhos que usaram à base TIMIT, é percebida a sua aplicação, unicamente, no reconhecimento de fonemas. Desse modo, difere dos objetivos dessa tese que é o reconhecimento da fala. Segundo DENG; YU (2014) as bases de falas contínuas apresentam uma taxa de erro de reconhecimento de aproximadamente 10%. Analisando a Tabela 5.14, o maior erro encontrado nos experimentos foi de 8,36%. Esse experimento serviu para testar a robustez do modelo aos sotaques, independência do locutor e sentenças de diferentes tamanho.

Em geral, o modelo adaptativo apresentou duas dificuldades: a primeira diz respeito à taxa de reconhecimento da fala quando o sinal é amostrado com pouca intensidade de ruído, como por exemplo, a 20dB. Essa dificuldade ocorreu porque a máscara do sinal filtrou componentes confiáveis do sinal. A segunda está relacionada à precisão da rede SOM. Utilizar os mesmos parâmetros para bases distintas ocasionou um número alto de falsos-positivos. Apesar dessas dificuldades, o modelo apresentou robustez ao ruído, sotaque e independência de locutor.

6

Conclusão

A reconstrução de características ausentes propicia uma maior robustez ao ruído, para os reconhecedores automáticos da fala, quando expostos a ambientes diferentes do seu treinamento. Na última década, vários pesquisadores exploraram a reconstrução de características ausentes que seja capaz de lidar com ruídos não estacionários.

Inicialmente, este trabalho replicou as principais técnicas para reconstrução de características ausentes da fala, utilizando a base Aurora 2, propondo a utilização do modelo acústico com aprendizagem não-supervisionada, tendo como finalidade o reconhecimento da fala com características ausentes. Nesse seguimento, a rede neural LARFSOM foi utilizada com as características provenientes do trabalho de KIM; STERN (2011).

Para resolver o problema de lentidão na etapa de competição do LARFSOM. Foi proposto o uso da técnica FNNS para encontrar os dois nodos vencedores, como discutido no Capítulo 4.

Como resultado, o modelo apresentou robustez ao ruído, sendo melhor que as principais técnicas descritas na literatura, quando treinada sem distinção de intensidade de ruídos. Entretanto, observamos que a técnica estava sensível a certos parâmetros e características utilizadas. Para identificar tais sensibilidades, utilizamos o método chamado "Hipercubo Latino". Esse método identificou os parâmetros da taxa de aprendizagem (ρ) e limiar de ativação (a_T) como os parâmetros mais sensíveis. Então, foi proposta uma equação para definir a taxa de aprendizagem, independente da base.

No segundo momento, foram feitas análises dos agrupamentos das características articulatórias. Para isso, utilizou-se as métricas estatística de precisão, *recall* e medida f. O modelo articulatório foi utilizado para verificar a influência da precisão na qualidade do agrupamento.

Por fim, adicionado as características articulatórias ao modelo, o modelo adaptativo foi treinado e testado utilizando as bases Aurora 2 e TIMIT. Para isso, através da rede SOM, protótipos articulatórios para identificar tais características foram criados.

Como resultado, o modelo adaptativo apresentou robustez ao ruído, sotaque e independência de locutor. Sendo treinado e testado em bases constituídas por números (Aurora 2), palavras e sentenças (TIMIT) de diferentes tamanho, objetivando a utilização em sistemas em tempo real.

O modelo adaptativo apresentou dificuldades na precisão da rede SOM e para reconstruir a fala em sinais com pouca intensidade de ruído. Isso ocorreu devido a máscara utilizada no sinal e por treinar a rede utilizando os mesmos parâmetros em bases distintas.

Como trabalho futuro, sugere-se utilizar novas características para definir a máscara do sinal, combinando-as com as características articulatórias que definem a maneira e o lugar da articulação. Para isso, sugere-se o uso de redes que considerem a relevância de cada característica no processo de reconhecimento da fala.

6.1 Contribuição para Ciência

1. Modelo adaptativo para reconhecimento da fala com reconstrução de características ausentes, aplicado a sistemas em tempo real.

2. Nova abordagem para identificar as características articulatórias utilizando a rede SOM.

3. Uso da aprendizagem não-supervisionada, utilizando bases distintas (com mesmo parâmetro), aplicado ao reconhecimento da fala com reconstrução de características ausentes.

4. Nova equação para encontrar a taxa de aprendizagem aplicada a rede LARFSOM.

5. Otimização, em relação ao tempo, da rede LARFSOM aplicando o FNNS.

6.2 Trabalhos Publicados e Submetidos

Os artigos listados, abaixo, foram publicados ou submetidos durante o desenvolvimento desta tese:

1. Viana, H.O e Araújo, A. F. R. **Adaptive speech model for missing-feature reconstruction**. International Conference on Digital Signal Processing, 2016.

2. Viana, H.O e Araújo, A. F. R. **New adaptive speech model for missing-feature speech recognition using acoustic and articulatory features**. IEEE Transactions on Audio, Speech, and Language Processing, 2017. - **Artigo Submetido**

Referências

ACERO; ALEJANDRO. **Acoustical and Environmental Robustness in Automatic Speech Recognition**. [S.l.]: Kluwer Academic Publishers, 1992.

AMITA, A.; BANSAL, P. Robust Features for Noisy Speech Recognition Using MFCC Computation from Magnitude Spectrum of Higher Order Autocorrelation Coefficients. **International Journal of Computer Applications**, [S.l.], v.10, p.36–38, 2010.

ANDRES, A. et al. A Pitch Based Noise Estimation Technique for Robust Speech Recognition with Missing Data. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2011. p.4808–4811.

ARAUJO, A. F. R.; COSTA, D. Local Adaptive Receptive Field Self-Organizing Map for Image Segmentation. In: WORKSHOP ON SELF-ORGANIZING NETWORKS. **Anais...** [S.l.: s.n.], 2007. p.1–8.

ARAUJO, A. F. R.; COSTA, D. Local Adaptive Receptive Field Self-Organizing Map for Image Color Segmentation. **Image and Vision Computing**, [S.l.], v.27, p.1229–1239, 2009.

BADIEZADEGAN, S.; ROSE, R. Mask Estimation in Non-Stationary Noise Environments for Missing Feature Based Robust Speech Recognition. **Speech Communication**, [S.l.], v.56, p.26–30, 2010.

BARKER, J.; COOKE, M.; ELLIS, D. Decoding Speech in the Presence of Other Sources. **Speech Communication**, [S.l.], v.45, p.5–25, 2005.

BARKER, J. et al. Soft Decisions in Missing Data Techniques for Robust Automatic Speech Recognition. In: INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING. **Anais...** [S.l.: s.n.], 2000. p.373–376.

BAUM, L.; PETRIE, T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. **Annals of Mathematical Statistics**, [S.l.], v.37, p.1554–1563, 1966.

BEHLAU, M. **Avaliação e Tratamento das Disfonias**. [S.l.]: Editora Lovise, 1995.

BISHOP, M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer-Verlag, 2006.

BISOL, L. **Introdução a Estudos de Fonologia do Português Brasileiro**. [S.l.]: Edipucrs, 2001.

BISOL, L.; BRESCANCINI, C. **Fonologia e Variação**: recortes do português brasileiro. [S.l.]: EDIPUCRS, 2002.

BLINN, J. What's the Deal With the DCT? **Computer Graphics and Applications**, [S.l.], v.13, p.78–83, 1993.

BOLL, S. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. **International Conference on Acoustics, Speech, and Signal Processing**, [S.l.], v.27, p.30–36, 1979.

BOURLARD, H.; MORGAN, N. **Connectionist Speech Recognition A Hybrid Approach**. [S.l.]: Kluwer Academic Publishers, 1994.

BREIMAN, L. Bagging Predictors. **Machine Learning**, [S.l.], v.24, p.123–140, 1996.

CHEUNG, E.; CONSTANTINIDES, A. Fast Nearest Neighbour Search Algorithms for Self-Organising Map and Vector Quantisation. **Signals, Systems and Computers**, [S.l.], v.2, p.946–950, 1993.

CLOUSE, D. et al. Time-Delay Neural Networks: representation and induction of finite-state machines. **IEEE Transaction on Neural Networks and Learning Systems**, [S.l.], v.8, p.1065–1070, 1997.

COOKE, M. et al. Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data. **Speech Communication**, [S.l.], v.34, p.267–285, 2001.

DAUBECHIES, I. **Ten Lectures on Wavelets**. [S.l.]: Society for Industrial and Applied Mathematics, 1992.

DAVIS, K.; BIDDULPH, R.; BALASHEK, S. Automatic Recognition of Spoken Digits. **Journal of the Acoustical Society of America**, [S.l.], v.24, p.637–642, 1952.

DELLER, J.; HANSEN, J.; PROAKIS, J. **Discrete-Time Processing of Speech Signals**. [S.l.]: Institute of Electrical and Electronics Engineers, 2000.

DEMPSTER, A.; LAIRD, N.; RUBIN, D. **Maximum Likelihood from Incomplete Data Via the EM Algorithm**. [S.l.]: Royal Statistical Society, 1977.

DENG, L.; YU, D. Deep Learning: methods and applications. **Foundations and Trends in Signal Processing**, [S.l.], v.7, p.197–387, 2014.

DHANANJAYA, N.; YEGNANARAYANA, B.; SURYAKANTH, V. G. Acoustic-Phonetic Information from Excitation Source for Refining Manner Hypotheses of a Phone Recognizer. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2011. p.5252–5255.

FREUND, Y.; SCHAPIRE, R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **European Conference on Computational Learning Theory**, [S.l.], v.55, p.119–139, 1995.

FRITZKE, B. A Growing Neural Gas Network Learns Topologies. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 7. **Anais...** [S.l.: s.n.], 1995. p.625–632.

FRY, D. B. Theoretical Aspects of Mechanical Speech Recognition. **Journal of the British Institution of Radio Engineers**, [S.l.], v.19, p.211–218, 1959.

GALES, M.; YOUNG, J. Robust Continuous Speech Recognition using Parallel Model Combination. **IEEE Transactions on Speech and Audio Processing**, [S.l.], v.4, p.352–359, 1996.

GAO, Y.; SU, H. **Controlling a Weighting Filter Based on the Spectral Content of a Speech Signal**. [S.l.]: Google Patents, 2003.

GAROFOLO, J. et al. **DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus**. [S.l.]: National Institute of Standards and Technology, 1990.

GIBBONS, J.; CHAKRABORTI, S. **Nonparametric Statistical Inference**. [S.l.]: Boca Raton, 2011.

GONZALEZ, A. et al. MMSE-Based Missing-Feature Reconstruction With Temporal Modeling for Robust Speech Recognition. **IEE Transaction on Audio, Speech, and Language Processing**, [S.l.], v.21, p.624–635, 2013.

GONZALEZ, R.; WOODS, R. **Digital Image Processing**. [S.l.]: Prentice Hall, 2008.

GRAVES, A.; MOHAMED, A.; HINTON, G. Speech Recognition with Deep Recurrent Neural Networks. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2013. p.6645–6649.

HARDING, P.; MILNER, B. Speech Enhancement by Reconstruction from Cleaned Acoustic Features. In: INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION. **Anais...** [S.l.: s.n.], 2011. p.1189–1192.

HAYKIN, S. **Redes Neurais Princípios e Práticas**. [S.l.]: BOOKMAN, 2001.

HERMANSKY, H.; ELLIS, D.; SHARMA, S. Tandem Connectionist Feature Stream Extraction for Conventional HMM Systems. **International Conference on Acoustics, Speech, and Signal Processing**, [S.l.], v.3, p.1635–1638, 2000.

HERMANSKY, H.; MORGAN, N. RASTA Processing of Speech. **IEEE Transactions on Speech and Audio Processing**, [S.l.], v.2, p.578–589, 1994.

HESS, W. **Pitch Determination of Speech Signals**: algorithms and devices. [S.l.]: Springer Series, 1983.

HORA, D.; COLLISCHONN, G. **Teoria Lingüística**: fonologia e outros temas. [S.l.]: Editora Universitária, 2003.

HORI, T.; KUBO, Y.; NAKAMURA, A. Real-Time One-Pass Decoding with Recurrent Neural Network Language Model for Speech Recognition. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2014. p.6364–6368.

HOSSAN, M.; MEMON, S.; GREGORY, A. A Novel Approach for MFCC Feature Extraction. In: INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING AND COMMUNICATION SYSTEMS. **Anais...** [S.l.: s.n.], 2010. p.1–5.

HUANG, F.; LEE, T. Pitch Estimation in Noisy Speech Using Accumulated Peak Spectrum and Sparse Estimation Technique. **IEEE Transactions on Audio, Speech, and Language Processing**, [S.l.], v.22, p.1833–1848, 2013.

ITU-T. **Digital Terminal Equipments – Coding of Analogue Signals by Pulse Code Modulation**. [S.l.]: Serie G: Transmission Systems and Media, Digital Systems and Networks, 2001.

JAIN A.; MURTY, M. F. P. Data Clustering: a review. **ACM Computing Surveys**, [S.l.], v.31, p.264–323, 1999.

JIAO, H.; ZHIMI, H.; CHAOCHENG, X. Pitch Detection Algorithm Based on NCCF and CAMDF. **International COnference on Computational Intelligence and Communication Networks**, [S.l.], v.4, p.661–664, 2012.

JOHNSTON, J. Transform Coding of Audio Signals Using Perceptual Noise Criteria. **IEEE Journal on Selected Areas in Communications**, [S.l.], v.6, p.314–323, 1988.

KIM, C.; STERN, R. Feature Extraction for Robust Speech Recognition Based on Maximizing the Sharpness of the Power Distribution and on Power Flooring. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2010. p.4574–4577.

KIM, W.; STERN, R. Band-Independent Mask Estimation for Missing-Feature Reconstruction in the Presence of Unknown Background Noise. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2006. p.1–6.

KIM, W.; STERN, R. Mask Classification for Missing-Feature Reconstruction for Robust Speech Recognition in Unknown Background Noise. **Speech Communication**, [S.l.], v.53, p.1–11, 2011.

KIRCHHOFF, K.; FINK, G.; SAGERER, G. Combining Acoustic and Articulatory Feature Information for Robust Speech Recognition. **Speech Communication**, [S.l.], v.37, p.303–319, 2002.

KOHONEN, T. **Self-Organizing Maps**. [S.l.]: Springer-Verlag, 1997.

KOLLIG, T.; KELLER, A. Efficient Multidimensional Sampling. In: **Anais...** [S.l.: s.n.], 2002. p.3210–3213.

KUHNE, M.; TOGNERI, R.; NORDHOLM, S. Mel-Spectrographic Mask Estimation for Missing Data Speech Recognition using Short-Time-Fourier-Transform Ratio Estimators. **International Conference on Acoustics, Speech and Signal Processing**, [S.l.], v.4, p.405–408, 2007.

LADEFOGED, P.; JOHNSON, K. **A Course in Phonetics**. [S.l.]: Cengage Learning, 2010.

LAMEL, L.; KASSEL, R.; SENEFF, S. Speech Database Development: design and analysis of the acoustic-phonetic corpus. **Speech Recognition Workshop**, [S.l.], v.2, p.161–170, 1986.

LEI, X.; LIN, H.; HEIGOLD, G. Deep Neural Networks with Auxiliary Gaussian Mixture Models for Real-Time Speech Recognition. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2013. p.7634–7638.

LI, B.; SIM, K. A Spectral Masking Approach to Noise-Robust Speech Recognition Using Deep Neural Networks. **IEEE Transactions on Audio, Speech, and Language Processing**, [S.l.], v.22, p.1296–1305, 2014.

LIPPMANN, R. Speech Recognition by Machines and Humans. **Speech Communication**, [S.l.], v.22, p.1–15, 1997.

MALLAT, S. **A Wavelet Tour of Signal Processing**. [S.l.]: Academic Press, 2009.

MANJUNATH, K.; RAO, K. Articulatory and Excitation Source Features for Speech Recognition in Read, Extempore and Conversation Modes. **International Journal of Speech Technology**, [S.l.], v.19, p.121–134, 2016.

MARSLAND, S.; SHAPIRO, J.; NEHMZOW, U. A Self-Organising Network That Grows when Required. **IEEE Transaction on Neural Networks and Learning Systems**, [S.l.], v.15, p.1041–1058, 2002.

MASJOODI, S.; VALI, M. Fuzzy Reconstruction of Cluster-Based Missing Features Method for Robust Speech Recognition. In: BIOMEDICAL ENGINEERING. **Anais...** [S.l.: s.n.], 2011. p.11–14.

MATEUS, M.; FALÉ, I.; FREITAS, J. **Fonética e Fonologia do Português**. [S.l.]: Universidade Aberta, 2005.

MEDDIS, R.; HEWITT, M. J. Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery. I: pitch identification. **Journal of Acoustical Society of America**, [S.l.], v.89, p.128–133, 1991.

MIKOLOV, T. et al. Recurrent Neural Network Based Language Model. In: INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION. **Anais...** [S.l.: s.n.], 2010. p.1045–1048.

MITCHELL, T. **Machine Learning**. [S.l.]: McGraw-Hill, Inc., 1997.

MITRA, V. et al. Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2012. p.4117–4120.

MITRA, V. et al. Articulatory Trajectories for Large-Vocabulary Speech Recognition. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2013. p.7145–7149.

MOORE, B. **An introduction to the psychology of hearing**. [S.l.]: Academic Press, 1989.

MORLET, J. et al. Wave Propagation and Sampling Theory. **Geophysics**, [S.l.], v.47, p.203–221, 1982.

NAESS, A.; LIVESCU, K.; PRABHAVALKAR, R. Articulatory Feature Classification using Nearest Neighbors. In: INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION. **Anais...** [S.l.: s.n.], 2011. p.2301–2304.

NAGATA, K.; KATO, Y. Spoken Digit Recognizer for Japanese Language. **Europe Research and Development**, [S.l.], v.12, p.336–342, 1963.

PAPCUN, J. et al. Inferring Articulation and Recognizing Gestures from Acoustics with a Neural Network Trained on X-Ray Microbeam Data. **Journal of the Acoustical Society of America**, [S.l.], v.92, p.688–700, 1992.

PATEL, I.; RAO, Y. Speech Recognition Using Hidden Markov Model with MFCC-Subband Technique. In: INTERNATIONAL CONFERENCE ON RECENT TRENDS IN INFORMATION, TELECOMMUNICATION AND COMPUTING. **Anais...** [S.l.: s.n.], 2010. p.168–172.

PEARCE, D.; HIRSCH, H.; EUROLAB, E. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions. In: INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING. **Anais...** [S.l.: s.n.], 2000. p.29–32.

POLUR, P.; MILLER, G. Experiments with Fast Fourier Transform, Linear Predictive and Cepstral Coefficients in Dysarthric Speech Recognition Algorithms using Hidden Markov Model. **Neural Systems and Rehabilitation Engineering**, [S.l.], v.4, p.558–561, 2005.

RABINER, L. et al. A Comparative Performance Study of Several Pitch Detection Algorithms. **IEEE Transactions on Acoustics, Speech and Signal Processing**, [S.l.], v.24, p.399–418, 1976.

RABINER, L.; JUANG, B. **Fundamentals of Speech Recognition**. [S.l.]: Prentice-Hall, Inc., 1993.

RABINER, L.; SCHAFER, R. **Digital Processing of Speech Signals**. [S.l.]: Prentice-Hall signal processing series, 1978.

RABINER, L.; SCHAFER, R. **Introduction to Digital Speech Processing**. [S.l.]: Foundations and Trends in Signal Processing Series, 2007.

RAJ, B.; SELTZER, M.; STERN, M. Reconstruction of Missing Features for Robust Speech Recognition. **Speech Communication**, [S.l.], v.43, p.275–296, 2004.

RICHARDS, H. et al. Vocal Tract Shape Trajectory Estimation using MLP Analysis-by-Synthesis. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 1997. p.1287–1290.

RêGO, R. L. M. E.; FERREIRA, P. H. M.; ARAúJO, A. F. R. Reconstructing Anatomical Structures with Growing Self-Reconstruction Maps. In: INTERNATIONAL CONFERENCE ON SYSTEMS, MAN, AND CYBERNETICS. **Anais...** [S.l.: s.n.], 2011. p.2976–2981.

SAKAY, T.; DOSHITA, S. The Phonetic Typewriter. **IRE Transactions on Audio**, [S.l.], v.11, p.140–152, 1962.

SAON, G.; SOLTAU, H. Boosting Systems for Large Vocabulary Continuous Speech Recognition. **Speech Communication**, [S.l.], v.54, p.212–218, 2012.

SCHWENK, H. Continuous Space Language Models. **Computer Speech & Language**, [S.l.], v.21, p.492–518, 2007.

SELTZER, M. **Automatic Detection Of Corrupt Spectrographic Features for Robust Speech Recognition**. [S.l.]: Carnegie Mellon University, 2000.

SELTZER, M.; RAJ, B.; STERN, R. A Bayesian Classifier for Spectrographic Mask Estimation for Missing Feature Speech Recognition. **Speech Communication**, [S.l.], v.43, p.379–393, 2004.

SENEFF, S. A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing. **Journal of Phonetics**, [S.l.], v.16, p.55–76, 1988.

SINGH, R.; STERN, R.; RAJ, B. **Model Compensation and Matched Condition Methods for Robust Speech Recognition**. [S.l.]: CRC Press, 2004.

SONG, Y.; PENG, X. Spectra Analysis of Sampling and Reconstructing Continuous Signal using Hamming Window Function. In: INTERNATIONAL CONFERENCE ON NATURAL COMPUTATION. **Anais...** [S.l.: s.n.], 2008. p.48–52.

STEVENS, S.; NEWMAN, E. A Scale for the Measurement of the Psychological Magnitude of Pitch. **Journal of the Acoustical Society of America**, [S.l.], v.8, p.185–190, 1937.

SUZUKI, J.; NAKATA, K. Recognition of Japanese Vowels Preliminary to the Recognition of Speech. **Journal Radio Research**, [S.l.], v.8, p.193–212, 1961.

TAN, L.; KARNJANADECHA, M. Pitch Detection Algorithm: autocorrelation method and amdf. **International Symposium on Communications and Information Technology**, [S.l.], v.3, p.551–556, 2003.

VIANA, H.; ARAUJO, A. F. R. Adaptive Speech Model for Missing-Feature Reconstruction. In: INTERNATIONAL CONFERENCE ON DIGITAL SIGNAL PROCESSING. **Anais...** [S.l.: s.n.], 2016. p.104–108.

VIANA, H.; MELLO, C. Speech Description Through MINERS: model invariant to noise and environment robust for speech. In: INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS. **Anais...** [S.l.: s.n.], 2014. p.489–494.

WAHID, R. et al. A Gaussian Mixture Models Approach to Human Heart Signal Verification using Different Feature Extraction Algorithms. **Computer Applications for Bio-Technology, Multimedia, and Ubiquitous City**, [S.l.], v.353, p.16–24, 2012.

WANG, W.; HARPER, P. The SuperARV Language Model: investigating the effectiveness of tightly integrating multiple knowledge sources. **Empirical Methods in Natural Language Processing**, [S.l.], v.10, p.238–247, 2002.