UNIVERSIDADE DO ESTADO DE SANTA CATARINA
CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT
ENGENHARIA ELÉTRICA

GUILHERME MARTIGNAGO ZILLI

SISTEMA DE RECONHECIMENTO DE FALA PARA PALAVRAS ISOLADAS COM VOCABULÁRIO RESTRITO E INDEPENDENTE DE LOCUTOR

JOINVILLE, SC
2012

GUILHERME MARTIGNAGO ZILLI

SISTEMA DE RECONHECIMENTO DE FALA PARA PALAVRAS ISOLADAS COM VOCABULÁRIO RESTRITO E INDEPENDENTE DE LOCUTOR

Trabalho de Conclusão apresentado ao Curso de Engenharia Elétrica do Centro de Ciências Tecnológicas, da Universidade do Estado de Santa Catarina, como requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Orientador: Aleksander Sade Paterno

JOINVILLE, SC

2012

GUILHERME MARTIGNAGO ZILLI

SISTEMA DE RECONHECIMENTO DE FALA PARA PALAVRAS ISOLADAS COM VOCABULÁRIO RESTRITO E INDEPENDENTE DE LOCUTOR

Trabalho de Conclusão apresentado ao Curso de Engenharia Elétrica do Centro de Ciências Tecnológicas, da Universidade do Estado de Santa Catarina, como requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Banca Examinadora

Orientador: _____
Prof. Dr. Aleksander Sade Paterno
Universidade do Estado de Santa Catarina

Membro: _____
Prof. Me. Raimundo Nonato Gonçalves Robert
Universidade do Estado de Santa Catarina

Membro: _____
Prof. Dr. Celso José Faria de Araújo
Universidade do Estado de Santa Catarina

Joinville, 06/07/2012

Aos meus pais.

AGRADECIMENTOS

Aos meus pais, por todo apoio, carinho e incentivo, fundamentais para que eu chegasse até onde estou.

Ao professor Aleksander S. Paterno, quem me orientou neste e em outros trabalhos, por todo o apoio.

Ao colega Lucas Negri, pela ajuda e pelas dicas que, sem dúvida, contribuíram para os resultados deste trabalho.

Aos amigos e colegas de curso, em especial ao Adriano, Dênis, Felipe Stein, Felipe Zimann e Ricardo, pelas conversas, trocas de experiência e apoio, durante estes quase cinco anos de graduação.

Aos professores do Departamento de Engenharia Elétrica da UDESC, pela importante contribuição em minha formação acadêmica. E aos professores André Leal, Ademir Nied, Marcos Fergutz e Yales R. Novaes, por importantes conselhos e palavras de incentivo.

Ao professor Adriano Bresolin, da UTFPR – Medianeira, pelas diversas dicas que me ajudaram a realizar este trabalho.

A todos os colegas que contribuíram com a gravação da base de dados utilizada no desenvolvimento do trabalho.

Ao Grupo PET Engenharia Elétrica, pela contribuição à minha formação e pelas oportunidades que tive ao participar do grupo, e aos colegas que trabalharam comigo durante o tempo que estive lá.

"Um súbito pensamento errôneo pode dar lugar a uma indagação frutífera que revela verdades de grande valor."
(Isaac Asimov)

RESUMO

ZILLI, Guilherme Martignago. **Sistema de Reconhecimento de Fala para Palavras Isoladas com Vocabulário Restrito e Independente de Locutor.** 2012. 82f. Trabalho de Conclusão de Curso (Bacharel em Engenharia Elétrica) – Universidade do Estado de Santa Catarina. Departamento de Engenharia Elétrica, Joinville, 2012.

Este trabalho descreve um sistema de reconhecimento automático de fala para palavras isoladas, com vocabulário restrito e independente de locutor. No trabalho são apresentados aspectos gerais da produção fisiológica da fala e sua modelagem digital; descrição das etapas de aquisição e pré-processamento do sinal de fala; discussão dos principais métodos utilizados na detecção e extração das palavras, apresentando ainda, uma proposta de modificação para um algoritmo já encontrado na literatura a fim de incrementar os resultados no processo de detecção de palavras; discussão das principais características usadas na representação paramétrica dos sinais de fala, dando atenção especial aos coeficientes mel-cepstrais; e discussão sobre as técnicas de classificação de padrões e uma breve explicação sobre o uso de redes neurais para esta finalidade. São apresentados, ainda neste trabalho, os experimentos realizados e os resultados obtidos na classificação dos comandos de voz, bem como a metodologia adotada para a criação do banco de dados utilizado no treinamento e validação do sistema.

Palavras-Chave: Reconhecimento de fala. Detecção de palavra. Coeficientes mel-cepstrais. Redes Neurais Artificias.

LISTA DE ABREVIATURAS

ADPCM	*Adaptive Differential Pulse Code Modulation*
ARPA	*Advanced Research Projects Agency*
DCT	*Discrete Cosine Transform*
DFT	*Discrete Fourier Transform*
DPCM	*Differential Pulse Code Modulation*
DTW	*Dynamic Time Warping*
FFT	*Fast Fourier Transform*
FIR	*Finite Impulse Response*
HMM	*Hidden Markov Model*
IIR	*Infinite Impulse Response*
LPC	*Linear Predictive Coding*
MFCC	*Mel-Frequency Cepstral Coefficients*
MLP	*Multi-Layer Perceptron*
PCM	*Pulse Code Modulation*
PMC	Perceptron Multi-Camadas
RNA	Redes Neurais Artificiais
RP	Reconhecimento de Padrões
RPROP	*Resillient Propagation*
SNR	*Signal to Noise Ratio*
STE	*Short Time Energy*
ZCR	*Zero Crossing Rate*

SUMÁRIO

1 INTRODUÇÃO 9
1.1 FUNDAMENTOS DO RECONHECIMENTO DE FALA 9
1.2 ESTADO DA ARTE 12
1.3 APLICAÇÕES 15
1.4 OBJETIVOS 16
1.5 ESTRUTURA DO TRABALHO 16
2 O SINAL DE FALA 18
2.1 PROCESSO DE PRODUÇÃO DA FALA NOS SERES HUMANOS 18
2.2 FONÉTICA ARTICULATÓRIA 20
2.3 SONS DA FALA E SUAS CARACTERÍSTICAS 21
2.4 MODELO DA PRODUÇÃO DA FALA 22
3 AQUISIÇÃO E PRÉ-PROCESSAMENTO DO SINAL DE FALA 25
3.1 AQUISIÇÃO DO SINAL DE FALA 25
3.2 PRÉ-PROCESSAMENTO 28
3.3 DETECÇÃO DA PALAVRA 31
3.3.1 Métodos Tradicionais 32
3.3.2 Métodos baseados na distribuição de probabilidade 35
3.3.3 Método proposto 36
4 EXTRAÇÃO DE CARACTERÍSTICAS 40
4.1 PRINCIPAIS REPRESENTAÇÕES PARAMÉTRICAS DO SINAL 43
4.1.1 Energia e amplitude média em um curto intervalo de tempo 43
4.1.2 Taxa de cruzamento por zero 44
4.1.3 Frequência fundamental ou *pitch* 44
4.1.4 Banco de filtros 45
4.1.5 Coeficientes de predição linear 46
4.2 COEFICIENTES MEL-CEPSTRAIS 49
5 RECONHECIMENTO DE PADRÕES 53
5.1 REDES NEURAIS ARTIFICIAIS 53
5.1.1 Neurônios 55
5.1.2 Arquitetura e treinamento de redes neurais artificiais 59
5.1.3 Rede Perceptron Multicamandas 60
5.1.4 O reconhecimento de padrões com redes neurais artificiais 62
6 EXPERIMENTOS 64
6.1 CONFIGURAÇÃO DOS EXPERIMENTOS 64
6.1.1 Base de dados 64
6.1.2 Pré-processamento e Extração de Características 64
6.1.3 Classificação de Padrões 66
6.2 EXPERIMENTO 1 66
6.3 EXPERIMENTO 2 67
6.4 EXPERIMENTO 3 67
7 RESULTADOS E DISCUSSÕES 69
7.1 EXPERIMENTO 1 69
7.2 EXPERIMENTO 2 70
7.3 EXPERIMENTO 3 71
8 CONCLUSÕES 74
9 REFERÊNCIAS 77
10 APÊNDICE A – ARTIGO ACEITO PARA PUBLICAÇÃO NO XIX CONGRESSO BRASILEIRO DE AUTOMÁTICA – CBA 2012 81

1 INTRODUÇÃO

A fala é uma forma natural de comunicação e nos permite transmitir informações, pensamentos, necessidades, e também, expressar sentimentos. A comunicação vocal tem ainda presença forte e muito importante na evolução humana, uma vez que, por um longo período, foi uma das únicas formas de transmitir conhecimento e experiências.

O fenômeno da fala é um processo tão espontâneo, que por vezes, sua complexidade passa despercebida. E é justamente essa facilidade, com que as pessoas estão habituadas, que motivou a busca por interfaces através de voz.

Por mais de cinco décadas, a interface homem-máquina através da fala tem motivado diversos trabalho de pesquisa em todo o mundo. O objetivo principal dessas interfaces é extrair as informações presentes no vocabulário e determinar seu conteúdo para utilização em outros sistemas.

1.1 FUNDAMENTOS DO RECONHECIMENTO DE FALA

O reconhecimento de fala consiste em um sistema responsável por extrair informações presentes no vocabulário e classificá-las de acordo com padrões existentes e/ou conhecidos, quanto ao seu conteúdo.

Os sistemas de reconhecimento de fala constituem-se então, basicamente de três subsistemas, sendo eles: de aquisição e processamento do sinal de fala; de extração de características; e de classificação/reconhecimento de padrões.

A aquisição e pré-processamento do sinal, consiste na etapa inicial do reconhecimento de fala. Nesta etapa ocorre a transformação do sinal de voz em um conjunto de dados digitais, apto a ser usado nas etapas posteriores.

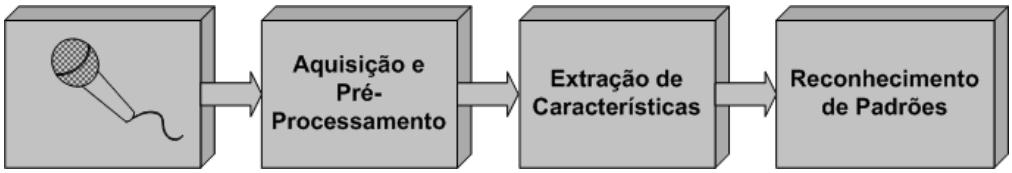
A extração de características permite reduzir o conjunto de dados dos padrões de áudio, resumindo-o e excluindo informações desnecessárias e/ou redundantes. A importância desta etapa deve-se ao fato de que a escolha correta das informações a serem extraídas pode incrementar o desempenho, tanto em termos de taxa de acerto como em relação ao tempo de processamento.

Finalmente, a etapa de classificação ou reconhecimento dos padrões, consiste basicamente na comparação dos dados, resultantes da etapa anterior, com

informações previamente conhecidas, relacionadas às palavras que o sistema é capaz de identificar.

A Figura 1 apresenta um diagrama de blocos básico de um sistema de reconhecimento de fala.

Figura 1 - Diagrama de blocos de um sistema de reconhecimento de fala.



Fonte: Autoria própria

De acordo com (RABINER e JUANG, 1993), um dos aspectos mais difíceis do reconhecimento de fala é sua característica interdisciplinar, pois envolve, entre diversas áreas, conhecimentos de:

Processamento de sinais – processo de extração de características relevantes do sinal, de uma maneira eficiente e robusta. Inclui também o pré-processamento e o pós-processamento para tornar o sistema robusto às características do ambiente.

Acústica – ciência que explica a relação entre o sinal de voz e o mecanismo fisiológico ligado a sua produção (trato vocal) e sua percepção (trato auditivo).

Reconhecimento de padrões – conjunto de algoritmos usados para agrupar dados e criar padrões baseados nas características extraídas, que serão usadas para comparação com os padrões a ser reconhecidos.

Linguística – relação entre sons (fonologia), palavras (sintaxe), significado da palavra (semântica) e o sentido. Inclui também a gramática e a análise do idioma.

Fisiologia – entendimento dos mecanismos fisiológicos de produção e percepção da fala, incluindo também o sistema nervoso central, responsável pelo seu processamento.

Ciência da computação – estudo dos algoritmos para implementação dos sistemas de reconhecimento de fala, em *software* ou em *hardware*.

Nas últimas décadas, muitos pesquisadores têm trabalhado e desenvolvido sistemas de reconhecimento de fala e de voz. O desenvolvimento e o aprimoramento dos algoritmos de extração de características e dos algoritmos de reconhecimento de padrões bem como a utilização de inteligência computacional, neste último, tem aproximado cada vez mais o desempenho dos sistemas de

reconhecimento de voz digitais do sistema humano. Todavia, não foi possível ainda desenvolver um sistema de reconhecimento da fala natural independente do ambiente, dos locutores (usuário) e de variáveis semânticas.

Desse modo, os sistemas desenvolvidos até o presente funcionam baseados em algumas restrições. Estas restrições permitem classificar os sistemas de reconhecimento de fala com relação a diversos aspectos, são eles: dependência de locutor, tamanho do vocabulário e estilo de fala.

1) Dependência de locutor: existem sistemas de reconhecimentos de fala dependentes e independentes de locutor (HUNT, 1997). Os sistemas independentes de locutor são sistemas capazes de reconhecer padrões provenientes de diversos falantes, mesmo que estes não tenham sido apresentados durante seu treinamento. Já os reconhecedores dependentes de locutor não têm habilidade para reconhecer comandos de usuários que não tenham sido utilizados no seu treinamento, possuindo um número de usuários muito limitado, porém, uma taxa de acerto superior ao sistema anterior. Uma classe especial de reconhecedores dependentes de locutor são os sistemas de reconhecimento de locutor, também chamado de sistemas de reconhecimento de voz. Estes sistemas têm por objetivo fazer a autenticação ou identificação de um usuário, podendo ser independente do conteúdo da palavra.

2) Tamanho do vocabulário: os vocabulários dos sistemas de reconhecimento de fala podem ser classificados em pequeno, médio e grande (NUNES, 1996). Os vocabulários considerados pequenos possuem entre 1 e 100 palavras, os vocabulários médios, entre 100 e 1000, e os vocabulários grandes, mais de 1000 palavras. O aumento no número de palavras pode reduzir o desempenho e a velocidade do sistema. Em função disso, vocabulários muito grandes passam a utilizar técnicas diferentes das utilizadas para vocabulários menores.

3) Estilo de fala: a classificação quanto ao estilo de fala pode distinguir o sistema entre reconhecimento de palavras isoladas, reconhecimento de palavras conectadas e reconhecimento de fala contínua (HUNT, 1997). O reconhecimento de palavras isoladas visa reconhecer o conteúdo de uma palavra falada individualmente. O reconhecimento de palavras conectadas, reconhecer o conteúdo de palavras faladas sequencialmente, mas de modo que cada palavra seja pronunciada claramente. Os sistemas de reconhecimento de fala contínua são sistemas que permitem reconhecer conteúdo de sentenças faladas de forma natural.

De acordo com (YOMA, 1993), “atualmente trabalha-se com sistemas que integram as técnicas de reconhecimento de fala desenvolvidas até agora, com modelos de linguagem natural, que introduzem conhecimento sintático e semântico do idioma, guiando o próprio reconhecimento de palavras e permitindo lidar com o significado de frases”.

1.2 ESTADO DA ARTE

O primeiro sistema de reconhecimento automático de fala, de acordo com (GOLD e MORGAN, 1999), foi um sistema construído pelos Laboratórios Bell, descrito em (DAVIS, BIDDULPH e BALASHEK, 1952). O sistema desenvolvido era capaz de reconhecer automaticamente dígitos, falados em velocidades normais, por um mesmo usuário, com taxa de acerto entre 97% e 99%. O funcionamento do sistema era baseado na divisão do sinal em duas bandas de frequência, uma superior, outra inferior a 900 Hz, cada uma contendo o primeiro e o segundo formante – formantes são frequências onde ocorrem picos de energia no espectro do sinal de fala –, respectivamente. O valor absoluto das frequências dos dois formantes e suas variações relativas eram usadas na comparação com padrões gerados previamente. A comparação era feita através do coeficiente de correlação entre os padrões conhecidos e as características do termo a ser classificado.

Até a década de 60, dezenas de sistemas de reconhecimento de fala foram desenvolvidos. Conforme apresentado em (GOLD e MORGAN, 1999), a grande maioria para reconhecer dígitos ou vogais, sendo testados com poucos locutores, raramente excedendo 10 ou 15 locutores. O reconhecimento de padrões era, em sua maioria, baseado em análise espectral e as taxas de acerto variavam entre 80% e 99.8%, caindo para até 44% para os sistemas capazes de reconhecer mais que 100 palavras. A análise espectral era feita, até então, por meio de bancos de filtros.

A partir dos anos 60 foram inseridas técnicas de análise espectral como a transformada rápida de Fourier (*Fast Fourier Transform* – FFT) (COOLEY e TUKEY, 1965) e análise cepstral, de análise temporal, como codificação linear preditiva (*Linear Predictive Coding* – LPC), e métodos para reconhecimento de padrões determinísticos, como a distorção de tempo dinâmica (*Dynamic Time Warp* –DTW), e estatísticos, como os modelos ocultos de Markov (*Hidden Markov Model* – HMM), (GOLD e MORGAN, 1999). Essas técnicas impulsionaram o desenvolvimento dos sistemas de reconhecimento automático de fala nos anos seguintes.

Em (RABINER, LEVINSON, *et al.*, 1979), foi proposto um sistema de reconhecimento de palavras isoladas independente de locutor para 39 palavras, treinado com cerca de 100 usuários, cuja classificação era baseada nos coeficientes de predição linear e usando também, técnicas como a DTW.

Até a meados da década de 80, pouca coisa mudou na estrutura dos sistemas de reconhecimento de fala. Aumentavam apenas a quantidade de dados e a dificuldade das tarefas. Nenhuma outra grande técnica foi desenvolvida durante este período (GOLD e MORGAN, 1999).

Ainda de acordo com (GOLD e MORGAN, 1999), entre 1971 e 1976, a *Advanced Research Projects Agency* (ARPA), agência do governo norte-americano, fundou um grande projeto para desenvolver um sistema de reconhecimento de fala contínua, para 1000 palavras, com poucos falantes e sujeito a restrições quanto a perda do sentido das sentenças. O objetivo foi alcançado apenas pelo sistema desenvolvido por Bruce Lowerre, estudante da Carnegie Mellon University. Na década seguinte, um segundo programa foi fundado pela ARPA, com comandos destinados a manipulação de uma base de dados de informações navais. Em 1998, já haviam sido desenvolvidos sistemas capazes de reconhecer mais de 60.000 palavras em tempo real, sem a necessidade de treinamento do locutor.

Um passo importante no desenvolvimento de sistemas de reconhecimento de fala foi o ressurgimento das redes neurais artificiais. As RNA já haviam sido utilizadas nos anos de 1950, mas sem sucesso. No entanto, pouco se havia aplicado de redes neurais nestes sistemas antes dos anos 1980. A partir do trabalho de (RUMELHART , HINTON e WILLIAMS, 1986), que reacendeu as pesquisas em redes neurais artificiais, muitas aplicação de RNA em reconhecimento de fala começaram a surgir.

Segundo (TEBELSKIS, 1995), as primeiras aplicações consistiam em tarefas simples, como classificação dos sons em vocálicos ou não vocálicos ou classificação entre sons fricativos, nasais ou plosivos. O sucesso nestas aplicações motivou os pesquisadores a utilizar RNA em reconhecimento de fonemas e, em seguida, de palavras.

Em (PERETTA, LIMA, *et al.*, 2009) é apresentado sistema de reconhecimento de comandos de voz para movimentação e troca de cor de um objeto em ambiente virtual bidimensional. O sistema é baseado na extração de características utilizando

transformada *wavelet* e na classificação de padrões com redes neurais. A taxa de acerto do sistema foi superior a 70%.

A evolução dos sistemas de reconhecimento automático de fala, bem como os principais projetos e instituições no qual tais projetos foram desenvolvidos são descritos em (JUANG e RABINER, 2005), (GOLD e MORGAN, 1999) e (FURUI, 2005).

Ao longo das últimas décadas, entre as técnicas de extração de características em sistemas de reconhecimento de fala, muitas variantes de bancos de filtros, codificação linear preditiva e análise cepstral foram utilizadas. Mais recentemente, a maioria desses sistemas convergiu para o uso de análise cepstral derivada de modelo do sistema auditivo. Entre as mais utilizadas estão a análise mel-cepstral e o LPC, (GOLD e MORGAN, 1999).

Já na etapa de reconhecimento e classificação de padrões, a maioria dos sistemas de reconhecimento automático de fala usam técnicas como Modelos Ocultos de Markov e Redes Neurais Artificiais. Existem ainda sistemas desenvolvidos com ambas as técnicas, chamados sistemas híbridos. Alguns detalhes dos sistemas híbridos são mostrados em trabalhos como (BOURLARD e MORGAN, 1998) e (TRENTIN e GORI, 2001).

Quando se trata de sistemas de reconhecimento de palavras isoladas, foco deste trabalho, existem ainda uma série de trabalhos relacionados.

Em (RUNSTEIN e VIOLARO, 1996) é descrito um sistema de reconhecimento de palavras isoladas, com vocabulário de 32 palavras e treinado para 24 usuários. No trabalho foi avaliado o desempenho de diversos algoritmos de extração de características, como o LPC, coeficientes cepstrais, coeficientes mel-cepstrais, energia, entre outros. Os vetores de características foram utilizados como entradas em duas redes neurais artificiais, uma Perceptron multicamadas e outra baseada na topologia Kohonen-Grossberg. As redes foram treinadas com 2/3 dos usuários e validadas com 1/3. Foram feitas ainda avaliações dos sistemas quando sujeito a ruído. Os mesmos apresentaram uma taxa de acerto de até 96% para o melhor caso.

Alguns trabalhos usam técnicas diferentes das que foram comentadas anteriormente, como em (SABAH e AINON, 2009), onde é apresentado um sistema de reconhecimento de dígitos isolados, independente de locutor, que utiliza um sistema de inferência fuzzy neuro-adaptativo como técnica de classificação. A base

de dados utilizada no treinamento foi montada a partir de 21 usuários. As proporções utilizadas para cada etapa foram as mesmas utilizadas no trabalho citado anteriormente.

Uma abordagem diferente é apresentada por (SAMOUELIAN, 1996). O autor apresenta a aplicação de uma técnica alternativa às RNAs e HHMs nos sistemas de reconhecimento de fala. No trabalho é proposta a utilização de inferência indutiva para criação de regras de classificação dos padrões.

De maneira geral, nos sistemas de reconhecimento de palavras isoladas, pode-se ainda observar que, grande parte dos trabalhos baseiam-se em técnicas como extração de coeficientes cepstrais e mel-cepstrais para formar o vetor de características das amostras.

1.3 APLICAÇÕES

Os sistemas de reconhecimento automático de fala possuem uma aplicabilidade extensa. Em (RUDNICKY, HAUPTMANN e LEE, 1994) os autores citam exemplos de aplicações de reconhecimento de fala em interfaces com computadores pessoais, onde comandos de voz podem ser usados como atalhos ou ainda, para executar tarefas quando se está com as mãos ocupadas, como por exemplo, alterar o estilo de fonte, enquanto o texto é digitado.

Outra linha de aplicação dos sistemas de reconhecimento de voz é no desenvolvimento de tecnologias assistivas. Portadores de deficiência motora devido à paralisia cerebral possuem uma grande dificuldade em se locomover e controlar eletrodomésticos. Neste sentido, em (SUK e KOJIM, 2008) é proposto um sistema de reconhecimento de fala aplicado à automação residencial inteligente e a uma cadeira de rodas. O reconhecimento de comandos de voz aplicado a uma cadeira de rodas se torna muito interessante nos casos de deficiências severas, como tetraplegia, onde o usuário não consegue utilizar interfaces como *joystick*. O reconhecimento de voz embarcado em cadeiras de rodas é apresentado também por (NETO, CASTRO e FELIX, 2010).

Já a aplicação dos sistemas de reconhecimento de voz aplicados à automação residencial é útil não apenas aos portadores de deficiências motoras, mas também aos portadores de deficiências visuais. Além disso, este sistema pode prover a todos os usuários, independente de deficiência, facilidades na realização de tarefas cotidianas e melhoria na qualidade de vida.

Existem também sistemas de reconhecimento de fala aplicados a sistemas de telefonia (NUNES, 1996) onde é possível fazer consulta a banco de dados, realizar operações e transações, ou ainda, a própria discagem.

1.4 OBJETIVOS

O objetivo principal deste trabalho é desenvolver um sistema de reconhecimento de fala para palavras isoladas, com vocabulário restrito e pequeno, independente de locutor e baseado em redes neurais artificiais.

O sistema desenvolvido deve ser capaz de reconhecer os seguintes comandos de fala: dígitos (zero, um,... , nove), comandos de direção (esquerda, direita, frente e atrás), comandos de ação (sobe, desce, liga, desliga, acende, apaga, abre e fecha) e ainda, as palavras casa, carro, luz e porta.

O desenvolvimento do trabalho seguiu as etapas descritas abaixo:

- Estudo e implementação de algoritmos para a aquisição e pré-processamento do sinal de fala;
- Estudo e implementação dos algoritmos de extração de características.
- Estudo e desenvolvimento de redes neurais artificiais aplicadas ao reconhecimento e classificação de padrões.
- Criação de um banco de dados com gravações de áudio com as palavras selecionadas, quando pronunciadas por diversos locutores, com o objetivo de realizar o treinamento e a validação do sistema desenvolvido.
- Avaliação do desempenho e analise dos resultados do sistema desenvolvido, utilizando os algoritmos implementados e a base de dados criada.

1.5 ESTRUTURA DO TRABALHO

Este trabalho está dividido da seguinte forma: na Seção 2 são apresentados os principais aspectos da produção da fala pelos seres humanos, as características dos sons produzidos e o modelo digital para produção dos sinais de fala; na Seção 3 são discutidos os processos de aquisição e pré-processamento dos sinais de fala e os algoritmos para detecção de palavras, é proposto ainda uma modificação à um algoritmo de detecção de palavras já existente na literatura; na Seção 4 são

apresentados os principais métodos de extração de características, dando atenção especial aos coeficientes mel-cepstrais; na Seção 5 são apresentados aspectos gerais do reconhecimento/classificação de padrões e da utilização de redes neurais artificiais com esta finalidade; a descrição dos experimentos e a apresentação dos resultados obtidos são mostrados nas Seções 6 e 7, respectivamente; finalmente, na Seção 8 são apresentadas as conclusões, considerações finais e propostas de trabalhos futuros.

2 O SINAL DE FALA

Nesta seção serão discutidos os principais aspectos fisiológicos e mecanismos de produção da fala pelos seres humanos. A partir de aproximações do modelo de produção da fala, será apresentada uma abordagem aproximada para um modelo digital de produção da fala.

A primeira etapa na produção do sinal de fala consiste na formulação da mensagem pelo falante. A partir da necessidade ou desejo de se transmitir uma informação ou ideia, o falante gera, em sua mente, uma mensagem que lhe permite expressar-se. Feito isso, o falante deve codificar sua mensagem em uma linguagem, e em seguida, através de ações neuromusculares, gerar vibrações nas suas cordas vocais que permitirão produzir uma sequência de sinais acústicos correspondendo à sua mensagem.

O processo de produção do sinal de fala humana pode ser estudado em diferentes aspectos, são eles: acústico, articulatório e auditivo.

O estudo do aspecto acústico está relacionado com a produção física do som, observando as grandezas como intensidade, frequências, duração, entre outras, e como tais grandezas permitem fazer a diferenciação entre os sons.

O aspecto articulatório estuda como são usados os órgãos do trato vocal (pulmão, laringe, boca, língua, etc.) e como eles interagem entre si para geração do som.

O aspecto auditivo (ou perceptual) está relacionado com o processo psicológico de produção e percepção do sinal de fala. Este aspecto estuda a interação entre o cérebro e o sistema auditivo, levando em conta, por exemplo, como o cérebro decodifica a informação acústica em uma mensagem, ou codifica a mensagem em uma informação acústica.

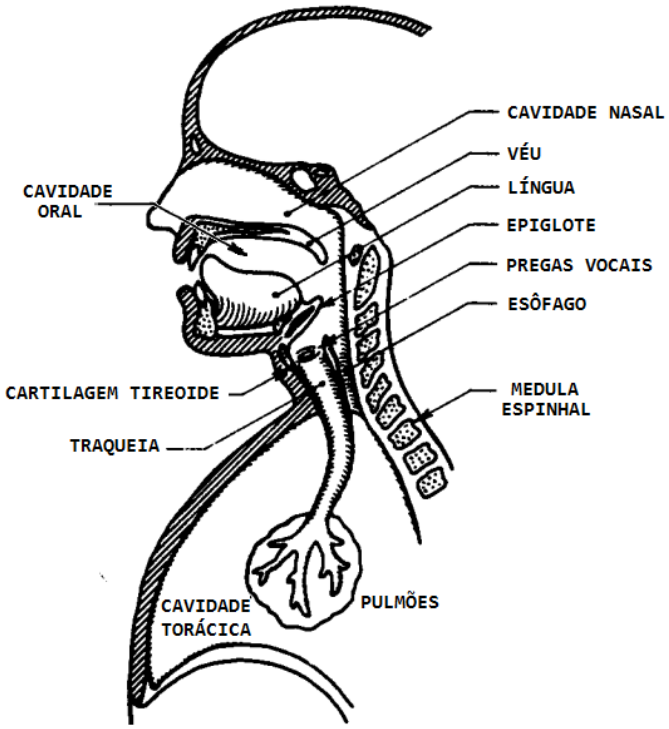
As ciências que estudam tais aspectos são chamadas, respectivamente, por fonética acústica, fonética articulatória e fonética auditiva.

2.1 PROCESSO DE PRODUÇÃO DA FALA NOS SERES HUMANOS

O ser humano não possui um órgão ou sistema exclusivo – ou com a função primária – para a produção da fala. O sinal de fala é produzido através da interação entre órgãos do sistema respiratório e digestivo. Os órgãos que participam da

produção da fala compõem o aparelho fonador. A Figura 2 apresenta uma vista esquemática do aparelho fonador humano.

Figura 2 - Diagrama do aparelho fonador.



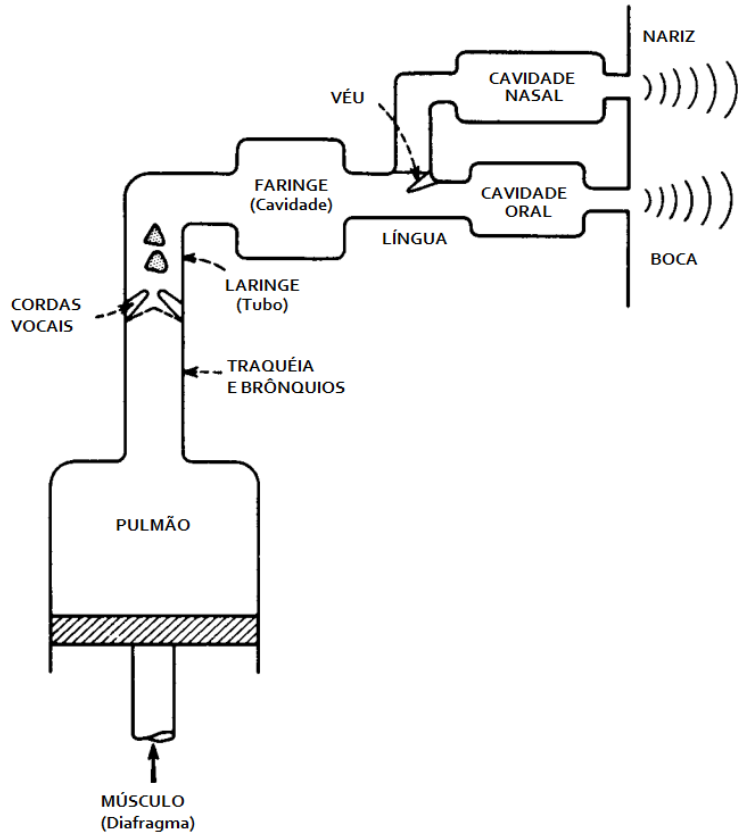
Fonte: adaptado de: (RABINER e JUANG, 1993)

O trato vocal, que tem início nas cordas vocais e termina nos lábio, consiste na faringe (conexão entre a cavidade oral e o esôfago) e na cavidade oral, tendo em média 17 cm para homens. O trato nasal inicia no véu e segue até as narinas. O véu funciona como uma válvula. Quando está aberta, o trato nasal e vocal são acoplados – o que permite produzir os sons nasais. Uma visão esquemática do mecanismo fisiológico de geração da fala é mostrada na Figura 3.

A produção da fala (ou fonação) ocorre sempre durante a expiração. O ar entra nos pulmões durante a inspiração e é expelido através da traqueia, passando pelas pregas vocais. As pregas (ou cordas) vocais podem estar abertas ou fechadas. Quando as cordas vocais estão fechadas, a passagem do ar é forçada, produzindo vibrações nas mesmas. Quando as cordas vocais estão abertas, o ar passa com facilidade por elas, causando pouca ou nenhuma vibração. Os sons produzidos quando as cordas vocais encontram-se fechadas são chamados de sons vozeados ou vocálicos, enquanto os sons produzidos quando elas estão abertas são

chamados de sons não vozeados ou não vocálicos. As transições entre abertura e fechamento das cordas vocais geram os sons transientes.

Figura 3 - Representação esquemática do mecanismo fisiológico de produção da fala.



Fonte: adaptado de (RABINER e JUANG, 1993).

O fluxo de ar no trato vocal é fatiado em pulsos quase periódicos que são modulados através da passagem pela faringe, pela cavidade oral e, possivelmente, pela cavidade nasal. Os diferentes sons produzidos durante a fala dependem da configuração das estruturas do trato vocal (posição da língua, lábios, véu, etc.).

A fonética articulatória explica como os sons são produzidos a partir das configurações do trato vocal.

2.2 FONÉTICA ARTICULATÓRIA

As estruturas que compõem o trato vocal e contribuem para a produção do som são chamadas de articuladores. Os principais articuladores, descritos em (SILVA, 2007), são apresentados a seguir:

Laringe: é o tubo que liga os pulmões à boca e ao nariz. Situa-se na parte superior da traqueia e é constituída principalmente de cartilagens e músculos. É na laringe que se encontram as cordas vocais e é também onde a voz é produzida.

Epiglote: localizada acima da laringe e das cordas vocais, é uma cartilagem, que se fecha a medida que o ar passa pelo trato vocal. Tem por função, também, evitar que alimentos ou água entrem pela laringe.

Faringe: é onde se localiza a raiz da língua. Faz a conexão entre a cavidade bucal e o esófago, e entre as fossas nasais e a laringe. É bastante flexível e, por este motivo, pode assumir diferentes formas e tamanhos.

Palato mole: também chamado de véu ou véu palatino, é uma estrutura muscular que controla a passagem de ar para a cavidade nasal. O véu é responsável por promover sons orais e nasais.

Palato duro: situa-se no céu da boca, próximo ao palato móvel. Embora seja um articulador fixo, a ação da língua, movendo-se em sua direção, pode produzir alguns sons (em geral, vocálicos).

Alvéolos: situado entre os dentes o palato duro, também é um articulador fixo e produz sons (em geral, consonantais) de acordo com a ação da língua.

Dentes: também um articulador fixo, produz sons devido a ação da língua.

Lábios: são articuladores móveis, e permitem produzir diversos movimentos, assim como, diversos sons.

Mandíbula: constitui o osso que forma o queixo e contribuí com os movimentos da língua e dos lábios.

Língua: é um dos articuladores mais importantes, pois apresenta grande flexibilidade e mobilidade, que lhe permite alcançar praticamente todas as outras estruturas do trato vocal. Por estas características, a língua participa na produção de praticamente todos os sons.

2.3 SONS DA FALA E SUAS CARACTERÍSTICAS

Os sons produzidos durante a fala são chamados de fonemas. Cada idioma apresenta um número distinto de fonemas em seus vocabulários. No inglês americano, por exemplo, existem 48 fonemas, incluindo 18 vogais ou combinações de vogais (ditongos), 21 consoantes, 4 vogais com sons consonantais, 4 sons silábicos e um fonema para representar a parada de um som vozeado (RABINER e JUANG, 1993).

É importante notar que os fonemas estão relacionados com os sons produzidos e são diferentes das letras.

A produção dos sons está diretamente relacionada com a posição e a atividade dos articuladores. De acordo com (OLIVEIRA) os sons linguísticos podem ser classificados quanto ao modo articulação. As classificações, para o idioma português-brasileiro são:

- Oclusiva ou plosiva: o fluxo de ar encontra uma interrupção total, seja pelo fechamento dos lábios, seja pela pressão da língua sob a arcada dentária ou sob o palato duro.
- Fricativa: há um estreitamento da passagem do ar, que resulta em um ruído semelhante ao de uma fricção.
- Africada: na realização dessas consoantes, temos a soma da oclusão e da fricção.
- Nasal: aquele som que, na sua realização, parte do ar sai pelo trato vocal e parte pelas fossas nasais.
- Lateral: a língua, ao tocar os alvéolos, obstrui a passagem do ar nas vias superiores, mas permite que o ar passe através das paredes laterais da boca.
- Vibrante: caracteriza-se pelo movimento vibratório e rápido da língua, provocando, assim, breves interrupções na corrente de ar.
- Tepe: ao contrário da vibrante, o tepe se caracteriza por apenas uma batida da ponta da língua nos alvéolos.
- Retroflexa: caracteriza-se pelo levantamento e encurvamento da ponta da língua em direção ao palato duro.

2.4 MODELO DA PRODUÇÃO DA FALA

A geração e propagação do som no trato vocal podem ser descritas por leis da física, em particular, pelas leis fundamentais de conservação de massa, conservação de momento e conservação de energia, bem como, pelas leis da termodinâmica e da fluidodinâmica, aplicadas à um fluído compressivo e de baixa viscosidade: o ar (RABINER e SCHAFER, 1978).Usando estes princípios, um

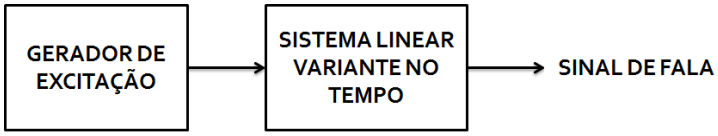
conjunto de equações diferenciais parciais poderá ser utilizado para descrever o movimento do ar no trato vocal.

O modelo detalhado do trato vocal deve considerar os seguintes efeitos: variação temporal da forma do trato vocal; perdas devido o aquecimento e o atrito viscoso nas paredes do trato vocal; suavidades das paredes do trato vocal; radiação do som nos lábios; acoplamento nasal; excitação do som no trato vocal.

Entretanto, não é do escopo deste trabalho resolver tais sistemas de equações diferenciais ou levar em conta todos estes efeitos. Uma modelagem mais detalhada do trato vocal pode ser encontrada em (RABINER e SCHAFER, 1978).

O modelo mais simplificado de produção da fala é apresentado na Figura 4

Figura 4 – Diagrama de blocos básico da produção da fala.



Fonte: adaptado de: (RABINER e SCHAFER, 1978).

O modelo da Figura 4 apresenta os dois blocos principais da produção da fala. O primeiro bloco representa a fonte de excitação do sinal. O segundo bloco apresenta-se como um sistema linear e variante no tempo. Este sistema tem como função, descrever o modelo do trato vocal, que age sobre o sinal de excitação. Tal modelo pode estar relacionado também à irradiação do sinal.

O sinal de excitação pode ser gerado no aparelho fonador através de três formas principais:

- o fluxo da ar sai dos pulmões e é modulado pela vibração das cordas vocais, resultando em uma excitação pulsada e quase periódica;
- o fluxo de ar se torna turbulento ao passar por uma constrição no trato vocal, resultando em uma excitação ruidosa;
- o fluxo de ar estabelece uma pressão em um ponto de fechamento total (constrição total) do trato vocal durante um pequeno intervalo de tempo. Ao remover a constrição, a pressão acumulada é liberada rapidamente, gerando uma excitação transitória.

O sistema linear que descreve o trato vocal está diretamente relacionado com a forma assumida pelo mesmo, ou seja, como a posição e atividade dos

articuladores. A relação entre a entrada e a saída deste sistema pode ser descrita por uma função de sistema $\mathrm{V}(z)$, como apresentado na Equação (1).

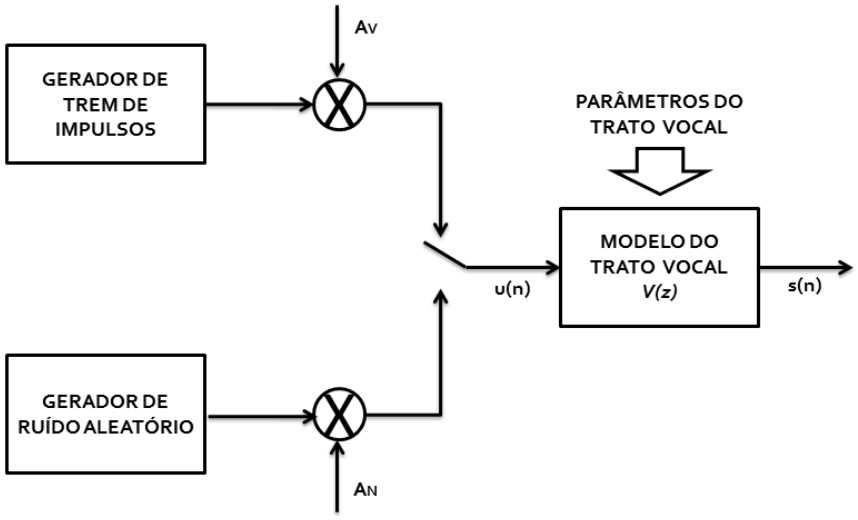
$$\mathrm{V}(z) = \frac{\mathrm{G}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \quad \textbf{(1)}$$

Onde G representa um ganho e os coeficientes a_k são os coeficientes do sistema linear.

Para produzir os diferentes sons, a característica desse sistema deve variar com o passar do tempo – por isto, sistema linear variante no tempo. No entanto, para sinais de fala, essas variações ocorrem lentamente. Para grande parte dos sons produzidos na fala assume-se que as propriedades (parâmetros) mantêm-se fixos por períodos entre 10 e 20 milissegundos (RABINER e SCHAFER, 1978).

Desta forma, um modelo um pouco mais completo da produção da fala pode ser elaborado, considerando um sistema linear variante no tempo, excitado por uma sequência periódica de pulsos (para sons vozeados) ou por um ruído aleatório (para sons não vozeados). Este modelo é mostrado na Figura 5.

Figura 5 – Modelo para produção da fala.



Fonte: adaptado de: (RABINER e SCHAFER, 1978)

3 AQUISIÇÃO E PRÉ-PROCESSAMENTO DO SINAL DE FALA

A primeira etapa no desenvolvimento do sistema de reconhecimento de fala consiste na aquisição e pré-processamento do sinal de áudio. O processo de aquisição está relacionado com a conversão das ondas de pressão sonora, gerada pelo trato vocal e captada pelo microfone, em um sinal digital. Em seguida, este sinal deve ser processado a fim de compensar algumas não linearidades, tanto do processo de produção da fala, quanto da própria aquisição. Outra etapa importante é a detecção de palavra. Esta etapa consiste na detecção dos limites de início e final da palavra para que a mesma possa ser extraída do restante do sinal adquirido.

Neste capítulo, são apresentadas, detalhadamente, as etapas de aquisição e pré-processamento do sinal de fala, bem como as principais técnicas de detecção de palavra. Será apresentado, também, um novo método de detecção de palavra, proposto neste trabalho.

3.1 AQUISIÇÃO DO SINAL DE FALA

O sinal de fala, quando produzido pelo aparelho fonador é propagado, na forma de ondas mecânicas – som – para o ambiente. Para que se possa utilizar este sinal, ele deve ser capturado e adequadamente formatado. A captura de sinais de áudio é, em geral, feita através do microfone, um transdutor que converte as ondas mecânicas do som em um sinal elétrico.

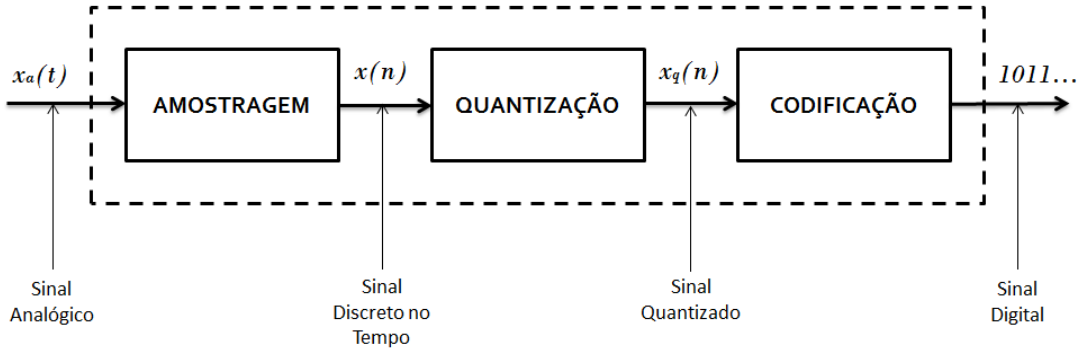
O sinal capturado pelo microfone se apresenta na forma de um sinal analógico – contínuo no tempo e em amplitude –, ou seja, o sinal pode assumir qualquer valor de amplitude durante todo intervalo de tempo. No entanto, para ser utilizado em sistemas digitais (como o sistema de reconhecimento de fala) é necessário que este sinal esteja em um formato adequado. Para tal, é necessário digitalizá-lo. Um sinal digital é um sinal discreto no tempo e em amplitude, isto é, existe apenas para certos valores discretos de tempo e sua amplitude pode assumir um número finito de valores.

A digitalização consiste na conversão de um sinal analógico para digital através dos processos de discretização temporal e da amplitude do sinal. Segundo (PROAKIS e MANOLAKIS, 1996), a conversão analógico-digital segue três etapas:

- Amostragem: converte um sinal contínuo no tempo em um sinal discreto no tempo. O sinal discreto no tempo é gerado através de amostras obtidas do sinal contínuo no tempo, em determinados instantes, periodicamente.
- Quantização: conversão de um sinal discreto no tempo e contínuo em amplitude em um sinal discreto no tempo e discreto em amplitude. A amplitude do sinal quantizado assume um valor selecionado dentro de um conjunto finito de valores possíveis.
- Codificação: no processo de codificação, cada valor assumido pelo sinal quantizado é codificado em uma sequência de bits.

O principal proposito do processo de digitalização é produzir uma representação dos dados amostrados com a maior relação sinal-ruído possível (PICONE, 1993). A Figura 6 apresenta o diagrama básico de um conversor de sinais analógico para digital.

Figura 6 - Diagrama de blocos do conversor analógico-digital.



Fonte: adaptado de (PROAKIS e MANOLAKIS, 1996).

O sinal digital pode ser caracterizado pelo período (ou frequência) de amostragem, pela resolução e pela codificação utilizada.

O período de amostragem é o tempo *(T_S)* entre duas amostras sucessivas. A frequência de amostragem equivale ao inverso do período *($F_S=1/T_S$)*. Mas a que frequência deve-se amostrar um sinal? De acordo com o Teorema da Amostragem, para que um sinal em banda base possa ser reconstruído a partir de suas amostras, é necessário que a amostragem seja feita a uma frequência pelo menos duas vezes maior que a maior frequência do sinal (LATHI, 2004), (PROAKIS e MANOLAKIS, 1996).

Como descrito anteriormente, na quantização o sinal assume um valor de um conjunto de valores possíveis. Estes ”valores possíveis” são chamados de níveis de quantização. O número de níveis de quantização equivale ao número de valores que o sinal pode assumir. A distância entre níveis de quantização sucessivos é chamada de passo de quantização ou resolução. Considerando um sinal com faixa dinâmica $(x_{max} - x_{min})$, onde x_{max} é o maior valor de amplitude assumido pelo sinal e x_{min}, o menor, e que o sinal seja quantizado em *L* níveis, o passo de quantização *(Δ)* é dado pela Equação (2).

$$\Delta = \frac{x_{max} - x_{min}}{L - 1} \quad \textbf{(2)}$$

Dessa forma, é possível perceber que quanto maior a quantidade de níveis de quantização, menor será o passo de quantização. Como consequência, melhor será a representação do sinal digitalizado, uma vez que o número de valores que o sinal pode assumir é maior. Além disso, o processo de quantização poder ser classificado em dois tipos: por truncamento e por arredondamento.

Na quantização por truncamento, desprezam-se as casas decimais menos significativa do valor quantizado a partir de uma casa decimal definida, como consequência, o resultado será sempre equivalente ao nível de quantização imediatamente abaixo do valor amostrado. Na quantização por arredondamento, o valor quantizado assumirá o valor do nível de quantização mais próximo.

A codificação é o processo, na conversão analógico-digital, que determina qual o número binário está relacionado com cada um dos níveis de quantização. Para cada nível de quantização há apenas um número binário associado, que o representa. As formas mais comuns de codificação são: PCM – *Pulse Code Modulation* –, DPCM – *Differential Pulse Code Modulation* – e ADPCM – *Adaptive Differential Pulse Code Modulation*.

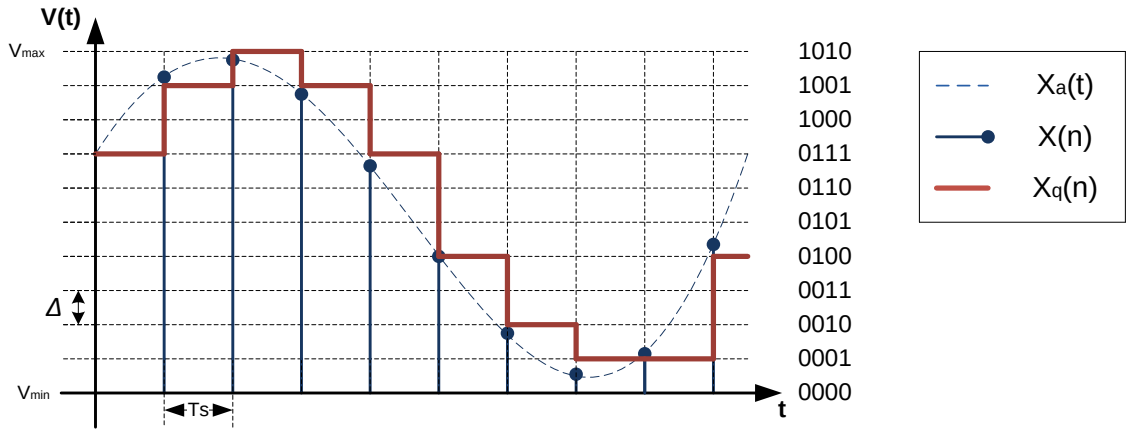
A codificação PCM é a forma mais simples de codificação e é a que será utilizada neste trabalho. Neste tipo de codificação, não há compressão do sinal e os valores resultantes da quantização são diretamente convertidos em uma palavra digital de *N* bits, onde $N + 1 = \log_2 L$ e *N* é um número inteiro.

A Figura 7 ilustra o processo completo de conversão analógico digital. O sinal analógico é representado pela linha tracejada – $x_a(t)$. A amostragem ocorre a cada

período T_S. O resultado da amostragem é exibido pela curva *x(n)*. O processo de quantização, neste caso, é por arredondamento. O resultado da quantização é exibido pela curva $x_q(n)$. No exemplo da Figura 7, a conversão analógico-digital resultaria na sequência de bits: 1001-1010-1001-0111-0100-0010-0001-0001-0100.

Uma abordagem mais detalhada da etapa de digitalização e dos processos envolvidos é apresentada em (PROAKIS e MANOLAKIS, 1996).

Figura 7 - Processo de digitalização.



Fonte: autoria própria

Após a digitalização, o sinal torna-se apto a ser manipulado e processado por sistemas e dispositivos digitais.

3.2 PRÉ-PROCESSAMENTO

O sinal de fala é inerentemente limitado em banda (frequência). Observa-se que o espectro do sinal de fala é atenuado em mais de 40 dB, com relação aos picos de maior intensidade, para frequências acima de 4 kHz, em sinais de fala vozeados. Para sinais de fala não vozeados, o espectro não sofre tamanha atenuação, mesmo para frequências de até 8 kHz, o que requereria taxas de amostragem maiores que 20 kHz (RABINER e SCHAFER, 1978).

Entretanto, boa parte da informação presente nos sinais de fala apresentam-se em frequências de até 3.5 kHz a 4 kHz. Os sistemas de telefonia, por exemplo, fazem a amostragem do sinal à 8 kHz, assumindo a maior frequência do sinal como sendo 4 kHz, e ainda assim, as informações transmitida pelo sistema são

perfeitamente inteligíveis. Dessa forma, pode ser conveniente limitar a banda do sinal a ser processado pelo sistema de reconhecimento de fala.

Além disso, o sinal de áudio, ao ser gravado em um dispositivo eletrônico, pode sofrer interferência conduzida ou irradiada de outras fontes, sejam sonoras ou eletromagnéticas. Os principais sinais de interferência, encontrados em gravações de áudio, são o ruído ambiente (sons de fundo, como emitido por aparelhos de ar-condicionado, carros, dispositivos de escritório – impressora, computador, etc.) e a interferência gerada pela rede elétrica.

Outros problemas podem ocorrer durante a etapa de aquisição do sinal. O microfone usado na captura do áudio e o processo de conversão A/D geralmente introduzem efeitos indesejados, como a perda de informações em baixas e altas frequências e distorções não lineares. Além disso, os conversores A/D possuem funções de transferências não lineares e acabam distorcendo o espectro do sinal (PICONE, 1993).

Outro aspecto importante, é que o espectro dos sons da fala tem uma perda total de -20 dB/década com o aumento da frequência. Esta perda é a combinação de uma perda de -40 dB/década causada pela fonte de excitação e o ganho de +20 dB/década devido à irradiação da boca, (MAGALHÃES, 2002).

Dessa forma, o pré-processamento tem por objetivo reduzir estes problemas. Este processo consiste em, quando uma amostra do sinal de áudio é digitalizada, executar uma filtragem digital deste sinal.

Inicialmente, faz-se a limitação da banda do sinal, desprezando componentes de alta frequência (maiores que 4 kHz, por exemplo) e de baixa frequência, em especial, a componente inserida pela interferência da rede elétrica, entre 50 Hz e 60 Hz. Para isso, pode-se utilizar um filtro de resposta ao impulso de duração finita – FIR – passa banda digital. A escolha das frequências de corte inferior e superior do filtro deve garantir a eliminação das frequências indesejadas.

Na sequência, o sinal foi filtrado pelo filtro de pré-ênfase. Segundo (PICONE, 1993) as duas principais vantagens da pré-ênfase são: compensar a atenuação natural, de aproximadamente 20 dB/década, devido às características fisiológicas do sistema de produção da fala e evidenciar componentes de frequência a partir de 1 kHz, faixa em que o sistema auditivo é mais sensível.

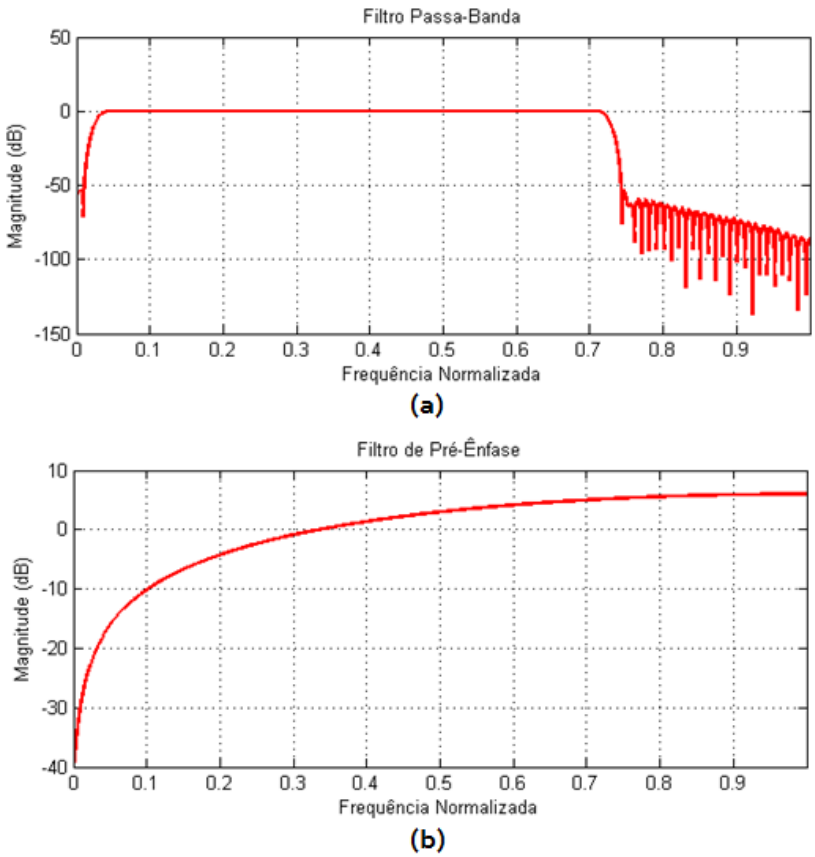
O filtro de pré-ênfase é, em geral, um filtro FIR com apenas um coeficiente, conhecido como filtro de pré-ênfase. A função de sistema do filtro de pré-ênfase é mostrada na Equação (3).

$$H_{pre}(z) = 1 + a_{pre} \cdot z^{-1} \quad \textbf{(3)}$$

O valor de a_{pre} pode variar entre [-0.4,-1], de acordo com (PICONE, 1993). Ficando em geral, em torno de -0.90 e -0.99. O filtro de pré-ênfase visa amplificar o espectro do sinal de fala em 20dB/década, sendo capaz de compensar as atenuações naturais do sinal.

Na Figura 8 são apresentadas as curvas de resposta em frequência para o (a) filtro de limitação de banda e para o (b) filtro de pré-ênfase. O filtro passa-faixa foi projetado para uma frequência de amostragem de 11025 Hz, com frequências de cortes inferior e superior de 150 Hz e 4000 Hz, respectivamente. O filtro de pré-ênfase foi projetado utilizando a_{pre} igual à -0.99.

Figura 8 - Resposta em frequência dos filtros de limitação de banda e de pré-ênfase.



Fonte: autoria própria.

Após a filtragem e pré-ênfase, o sinal é normalizado em amplitude e segue para a etapa de detecção de palavra.

3.3 DETECÇÃO DA PALAVRA

A etapa de detecção de palavras é chamada também de "*endpoint detection*". De acordo com (LAMEL, RABINER, *et al.*, 1981), o *endpoint detection* consiste na detecção dos pontos de início e final da pronúncia da palavra presente em uma gravação. Esta detecção permite extrair apenas o intervalo onde há sinal de fala, excluindo os intervalos em que há apenas ruído de fundo. Ainda segundo os autores, a extração da palavra é um processo muito importante, pois reduz o processamento do sinal e está diretamente relacionado à precisão do sistema de reconhecimento de fala.

O problema da distinção do sinal de fala do ruído de fundo não é trivial, exceto nos casos onde a gravação apresenta uma alta relação sinal-ruído (SNR – *Signal-to-Noise Ratio*), por exemplo, gravações feitas em câmaras anecóicas ou salas a prova de som. Nestes casos, os menores níveis de energia do sinal de fala ainda supera a energia do ruído de fundo, permitindo que a medida da energia seja suficiente para determinar os limites da palavra. No entanto, na maioria das aplicações práticas não se atingem as condições ideais para gravação (RABINER e SCHAFER, 1978).

As maiores dificuldades na detecção do início e final da pronúncia, de acordo com (RABINER e SCHAFER, 1978), estão em:

- Palavras começadas ou terminadas com fricativos fracos;
- Palavras começadas ou terminadas com plosivos fracos;
- Palavras terminadas com sons nasais;
- Fricativos vocálicos que se tornam não-vocálicos no final da palavra;
- Desvanecimento do som de uma vogal no final de uma pronúncia.

Outros problemas que costumam aparecer, quando se deseja separar o sinal de fala do ruído de fundo, pode ser estar relacionados a elementos não estacionários do sinal de ruído ambiente como, por exemplo, bater de portas, movimento de cadeiras, conversas, etc., ou ainda, a sons gerados pelo locutor,

como ruídos da boca, por exemplo, ao bater ou lamber os lábios e respirar (LAMEL, RABINER, *et al.*, 1981).

3.3.1 Métodos Tradicionais

Os dois métodos mais aceitos, e que são usados há muito tempo, são baseados na energia de um curto intervalo de tempo (*Short Time Energy – STE*), taxa de cruzamento por zero (*Zeros Crossing Rate – ZCR*) (estes métodos serão mais bem descritos em seções seguintes). O método da energia de um curto intervalo de tempo usa o fato de que, em um som vozeado, a energia do sinal é maior que para sons não vozeados e/ou para o ruído de fundo. Por outro lado, a taxa de cruzamento por zero apresenta valores quase cinco vezes maiores para sons não vozeados ou ruído ambiente. No entanto, tais métodos apresentam suas limitações, pois os ajustes dos limiares dependem dos dados existentes. (SAHA, CHAKRABORTY e SENAPATI, 2005).

Embora o sistema detecção de palavras possa operar exclusivamente com o método STE ou ZCR, é comum que estes dois métodos estejam associados.

Um exemplo de algoritmo utilizando a energia e a taxa de cruzamento por zero é apresentado em (RABINER e SAMBUR, 1975). As equações para o cálculo da energia do sinal de fala e da taxa de cruzamento por zero são definidas pelas Equação (4) e Equação (5), respectivamente. De acordo com o autor, o cálculo da energia é feito com o somatório do módulo da amplitude, ao invés do quadrado, para reduzir os esforços computacionais. Já a taxa de cruzamento por zero representa o número de alterações no sinal da amplitude do sinal de fala, o que equivale ao número de passagens pelo nível zero.

$$\mathrm{E}_n = \sum_{m=-N/2}^{N/2} |\mathrm{x(n+m)}| \tag{4}$$

$$\mathrm{Z}_n = \frac{1}{2} \cdot \sum_{m=-N/2}^{N/2} |sgn[\mathrm{x(m)}] - sgn[\mathrm{x(m-1)}]| \tag{5}$$

Onde, $\mathrm{x(n)}$ representa o sinal de fala, N o comprimento da janela onde o sinal será analisado e $sgn[\mathrm{x(m)}]$ representa a função sinal, definida pela Equação (6).

$$sgn[\mathrm{x(n)}] = \begin{cases} -1, & \mathrm{x(n)} < 0 \\ 1, & \mathrm{x(n)} \geq 0 \end{cases} \quad (6)$$

No algoritmo apresentado a magnitude média e a taxa de cruzamento por zero são calculadas a cada 10 ms (a uma taxa de 100 vezes/segundo). Além disso, assume-se que os primeiros 100 ms do intervalo de gravação não contém informações de fala (intervalo de silêncio). Esse pequeno intervalo é utilizado para caracterização do ruído de fundo – são calculados o desvio padrão a média da taxa de cruzamento por zero e os níveis de energia do ruído – para determinação dos limiares.

Os limiares são determinados por:

$$IZCT = \mathrm{MIN}(\mathrm{IF}, \overline{IZC} + 2\sigma_{IZC}) \quad (7)$$

$$I1 = 0.03 \cdot (\mathrm{IMX} - \mathrm{IMN}) + IMN \quad (8)$$

$$I2 = 4 \cdot IMN \quad (9)$$

$$ITL = \mathrm{MIN}(\mathrm{I1}, \mathrm{I2}) \quad (10)$$

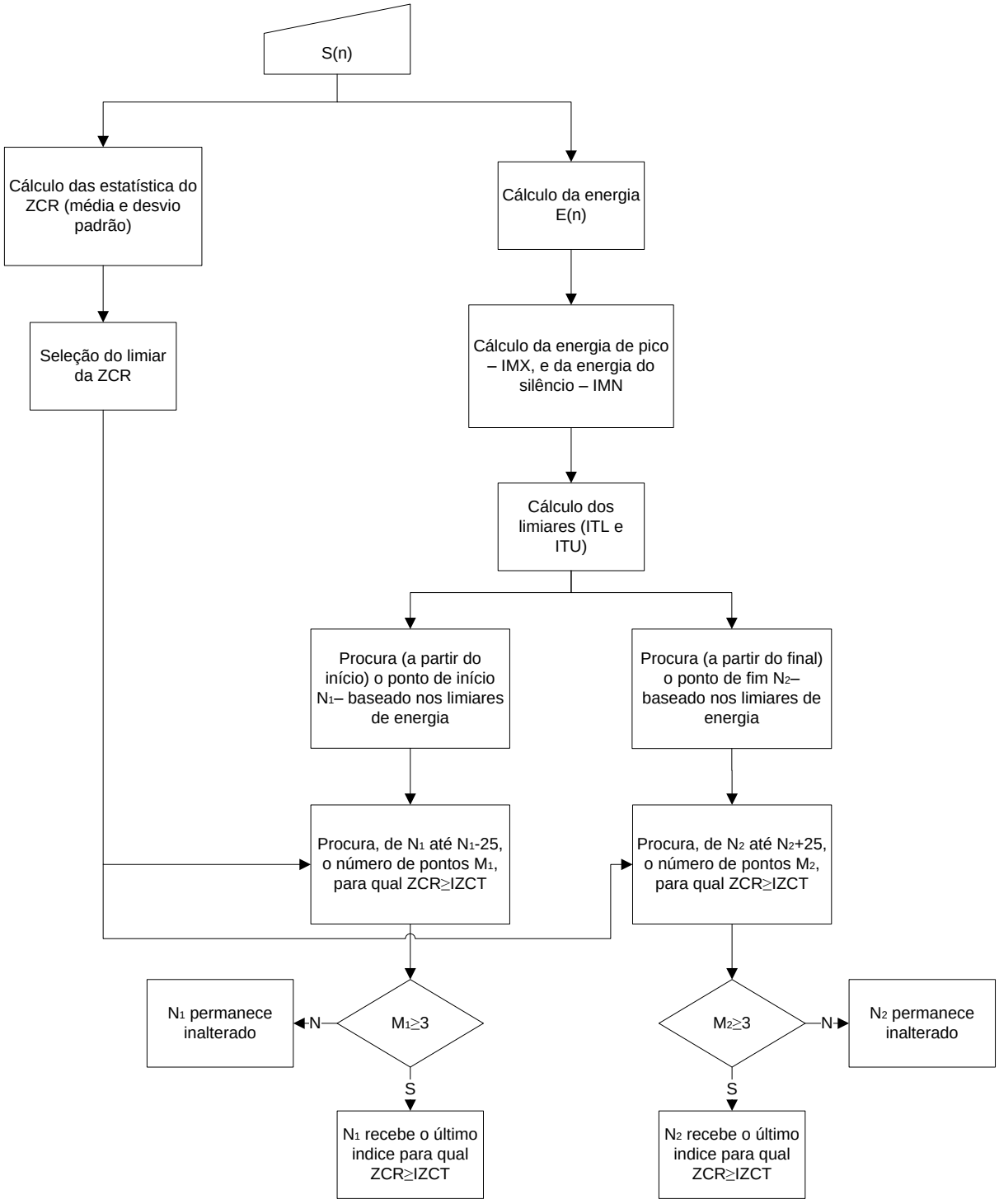
$$ITU = 5 \cdot \mathrm{ITL} \quad (11)$$

Onde, *IZCT* é o limiar da ZCR, *IF* é um limiar fixo (25 cruzamentos por zero por 10 ms), $\overline{IZC}$ e σ_{IZC} são a média e o desvio padrão da ZCR no intervalo de silêncio, *IMX* é o maior pico de energia do sinal de fala, *IMN* é a energia média do intervalo de silêncio e *ITL* e *ITU* são os limiares de energia inferior e superior, respectivamente.

O algoritmo pode ser explicado pelo fluxograma da Figura 9. Maiores detalhes são apresentados em (RABINER e SAMBUR, 1975).

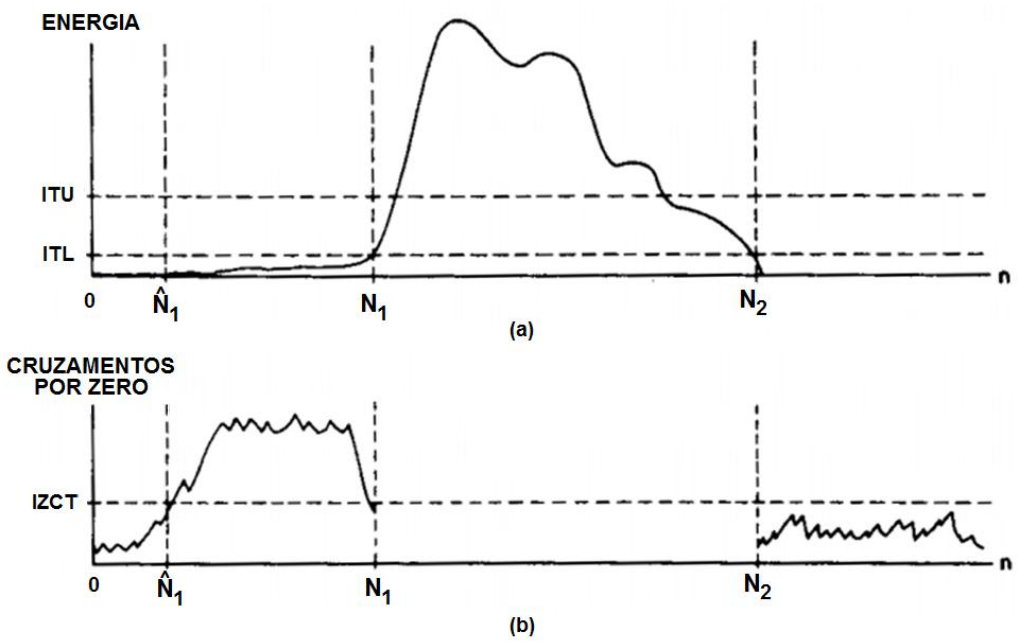
A Figura 10 apresenta um exemplo de medida de energia e taxa de cruzamento por zero para uma palavra que começa com um som fricativo forte. Na figura são apresentados os instantes em que o algoritmo detecta os limiares de STE e ZCR.

Figura 9 - Fluxograma do algoritmo de detecção de palavras proposto por (RABINER e SAMBUR, 1975).



Adaptado de: (RABINER e SAMBUR, 1975)

Figura 10 - Exemplo típico de medida de STE e ZCR para uma palavra começando com som um fricativo forte.



Adaptado de: (RABINER e SAMBUR, 1975)

3.3.2 Métodos baseados na distribuição de probabilidade

Outros métodos comuns para detecção de palavras consistem nos métodos baseados na distribuição de probabilidade do sinal. Estes algoritmos são baseados em propriedades estatísticas do ruído ambiente e em aspectos fisiológicos da produção da fala. Como principal vantagem, não necessita do cálculo de nenhum limiar (SAHA, CHAKRABORTY e SENAPATI, 2005) (KEERIO, MITRA, *et al.*, 2009).

Nos métodos baseados na distribuição de probabilidade, considera-se que o ruído possui distribuição de probabilidade, em geral, gaussiana.

No algoritmo apresentado em (SAHA, CHAKRABORTY e SENAPATI, 2005), por exemplo, assim como no algoritmo descrito anteriormente, considera-se que o início da gravação contenha apenas a ruído ambiente. Para este intervalo de silencio, são calculados a média (vide Eq. (12)) e o desvio padrão (Vide Eq. (13)) das amostras do sinal.

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x(n) \qquad (12)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n) - \mu)^2} \quad \textbf{(13)}$$

A parir destes parâmetros – que descrevem a função de densidade de probabilidade – o algoritmo percorre todas as amostras do sinal, checando a distância de Mahalanobis, definida pela Equação (14). Caso a distância *r* seja maior ou igual a três, a amostra é marcada como sinal de fala, caso contrário, a amostra é marcada como silêncio. Após este procedimento, a gravação é dividida em janelas de 10 ms. Para cada janela, compara-se o número de amostras marcadas como sinal de fala com o número de amostras marcadas como silêncio. As janelas que apresentarem um número de amostras marcadas como sinal de fala maior que o número de amostras marcadas como silêncio, são extraídas e concatenadas, formando uma nova sequência considerada como sinal de fala.

$$r = \frac{|x(n) - \mu|}{\sigma} \quad \textbf{(14)}$$

A principal desvantagem destes métodos consiste no fato de que, se a distribuição de probabilidade do sinal de ruído não for descrita razoavelmente bem pela distribuição de probabilidade escolhida, o algoritmo pode falhar completamente (KEERIO, MITRA, *et al.*, 2009).

3.3.3 Método proposto

O método escolhido para este trabalho foi baseado no método apresentado por (BRESOLIN, 2008) com algumas modificações.

Em seu trabalho, (BRESOLIN, 2008) propõem que, inicialmente, o sinal seja segmentado em janelas de 5 ms, e seja calculada a energia para cada segmento, formando um vetor de energia. Como existem muitos segmentos, grande parte deles representam apenas o sinal de ruído de fundo. Segundo o autor, a mediana do vetor energia representa, aproximadamente, a energia do sinal do ruído de fundo.

Dessa forma, o limiar de separação entre o sinal de fala e o ruído de fundo é dado pelo valor de três vezes a mediana do vetor energia. O valor de três vezes foi

definido empiricamente. Assim, quando a energia do sinal ultrapassasse o limiar de separação, o sinal seria considerado sinal de fala.

Neste trabalho propõe-se uma modificação no algoritmo apresentado por (BRESOLIN, 2008). A proposta de alteração foi motivada pelos problemas apresentados em (LAMEL, RABINER, *et al.*, 1981) e já discutidos neste trabalho, como por exemplo, a captura de sons gerados pelo locutor, como ruídos da boca, (i.e., ao bater ou lamber os lábios e respirar).

Estes sons apresentam, em geral, um pico elevado de energia, que ultrapassa o limiar de três medianas, podendo ser classificado como som de fala. No entanto, possuem uma duração curta.

A principal modificação proposta neste trabalho foi a consideração de um limiar temporal, ou seja, é necessário que a energia do segmento do sinal seja sempre maior que o limiar de detecção, durante um período equivalente a 10 janelas, para que seja considerado o início da palavra. De maneira análoga, para determinação do final da palavra é necessário que a energia do segmento do sinal seja sempre menor que o limiar de detecção, durante um período equivalente a 10 janelas.

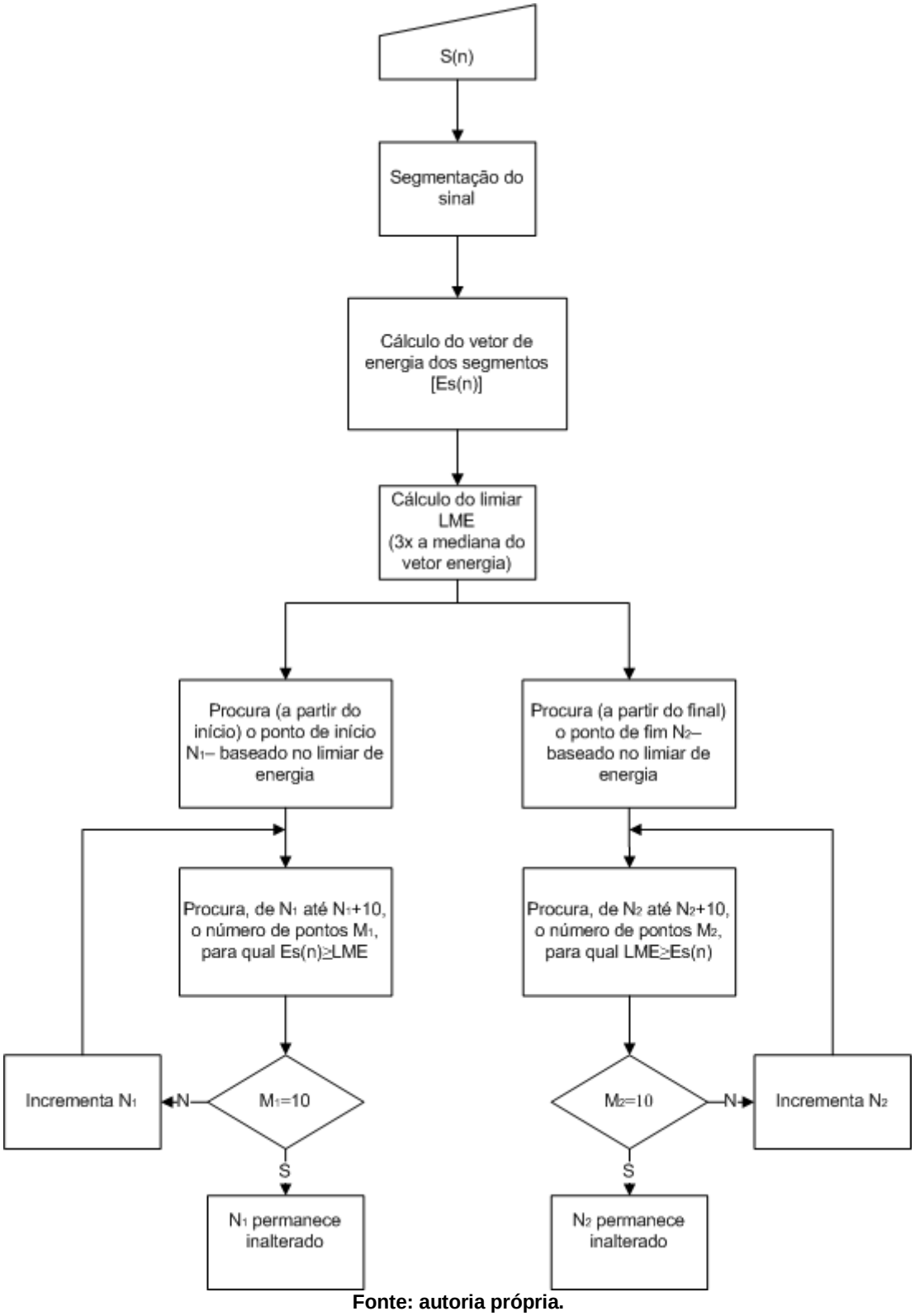
Além disso, neste trabalho foram utilizados janelas de tamanho aproximado de 3.6 ms e limiar de separação equivalente à quatro vezes o valor da mediana.

A Figura 11 apresenta o fluxograma do algoritmo de detecção de palavras apresentado neste trabalho.

Na Figura 12 é apresentada uma comparação entre os dois algoritmos. Na figura é mostrada a forma de onda – no domínio do tempo – do sinal correspondente à pronúncia da palavra “frente”. A linha pontilhada corresponde ao resultado obtido através do algoritmo original e a linha sólida, ao resultado do algoritmo proposto.

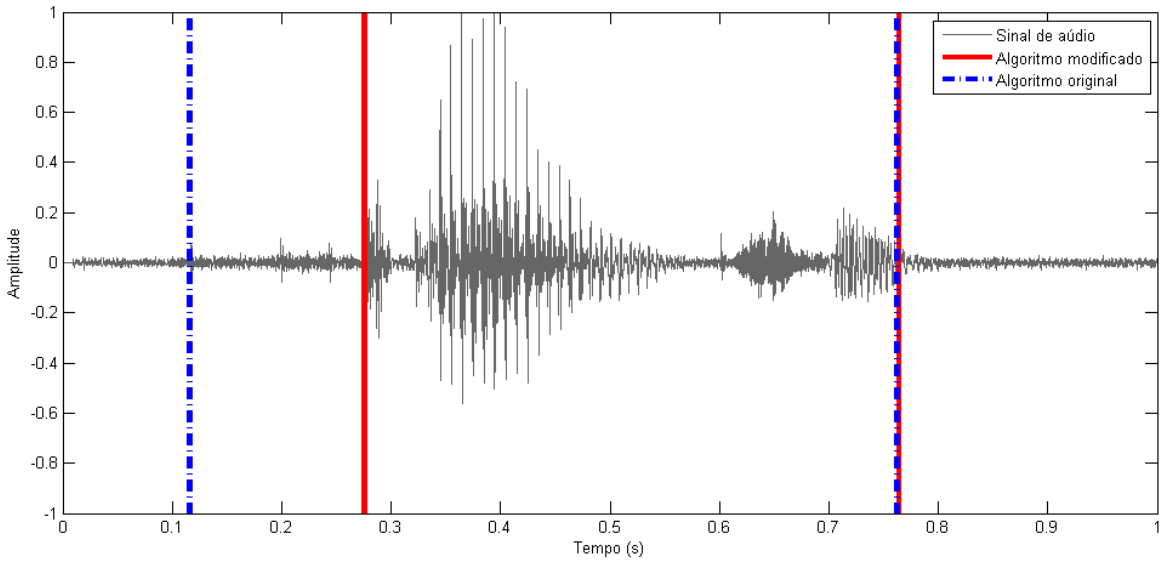
A comparação entre os resultados permite observar que o algoritmo proposto é capaz de desconsiderar picos de energia de curta duração, tornando-o mais robusto aos problemas apresentados em (LAMEL, RABINER, *et al.*, 1981). Além disso, percebe-se, visualmente, que os limiares encontrados pelo algoritmo proposto se ajustam melhor aos limiares reais da palavra.

Figura 11 - Fluxograma do algoritmo de detecção de palavras proposto neste trabalho.



Fonte: autoria própria.

Figura 12 - Comparação entre o resultado obtido pelo algoritmo proposto por (BRESOLIN, 2008) e o algoritmo pro proposto neste trabalho.



Fonte: autoria própria

4 EXTRAÇÃO DE CARACTERÍSTICAS

A aquisição e pré-processamento do sinal de áudio resultam em uma grande quantidade de dados. Neste trabalho, por exemplo, os dados são adquiridos a uma taxa de 11025 amostras/segundo, com resolução de 16 bits, o que equivale a uma taxa de geração de dados de 176.4kbits/s.

Ainda, segundo (BRESOLIN, 2008), dois sinais de voz (no domínio do tempo) de uma mesma palavra são diferentes, independente de a palavra ter sido pronunciada pela mesma pessoa ou não. Dessa forma, para fazer o reconhecimento de uma palavra é preciso encontrar determinadas características do sinal de voz que se repetem quando esta mesma palavra for pronunciada. Tais características descrevem um determinado padrão de voz, permitindo classificá-lo.

O objetivo da extração de características é descrever o sinal, através de um novo conjunto de dados, que apresentam as informações mais significativas deste sinal. Segundo (ZHENG, 2007), a extração de características consiste em uma transformação do sinal, visando, principalmente: destacar as informações mais significativas deste sinal; reduzir redundâncias das informações; e aumentar a robustez quanto ao ruído.

Em um aspecto prático, essa transformação consiste na projeção de um vetor, que representa o sinal, expresso por $\boldsymbol{s} = [s_1, s_2, s_3, \ldots, s_n]$, e com dimensão n, em um subespaço de dimensão p, resultando num outro vetor, também chamado de vetor de características, $\hat{\boldsymbol{s}} = [\hat{s}_1, \hat{s}_2, \hat{s}_3, \ldots, \hat{s}_p]$, onde, em geral, tem-se $p \leq n$.

Embora alguns autores tratem a extração de característica por este nome (GOLD e MORGAN, 1999), outros, como (PICONE, 1993), preferem chamar de modelagem de sinal. Segundo o autor, o termo “extração de característica” leva a crer que a quantidade de informações é reduzida – o que nem sempre acontece – e ainda, sugere que a utilização do termo implica que saibamos quais características são procuradas no sinal. Neste trabalho serão utilizados ambos os termos quando se tratar do conjunto de técnicas utilizadas para a representação paramétrica do sinal de fala.

Segundo (THEODORIDIS e KOUTROUMBAS, 2008), a extração da característica é uma etapa fundamental nos problemas de reconhecimento/classificação de padrões.

A importância desta etapa deve-se ao fato de que a escolha correta das informações a serem extraídas pode incrementar o desempenho, tanto em termos de resultados, como também, em relação ao tempo de processamento. Dessa forma, o maior problema no projeto de sistemas de extração de características consiste na seleção de quais características serão utilizadas e na determinação da quantidade de informações necessária e suficiente para descrever os dados originais.

No projeto de algoritmo para extração de características do sinal de fala, os principais esforços, se concentram, basicamente, em três aspectos (PICONE, 1993):

- Encontrar parâmetros que descrevam características importantes do sinal de fala. Preferencialmente estes parâmetros devem estar relacionados com o modelo usado para o trato vocal.
- A parametrização deve ser desejavelmente, robusta às variações do canal, do locutor e do transdutor.
- Devem apresentar parâmetros que capturam a dinâmica espectral do sinal, ou seja, as mudanças do espectro com o tempo.

Existe uma grande quantidade de representações paramétrica do sinal de fala, incluindo energia de um curto intervalo de tempo, taxa de cruzamento por zero, entre outros (RABINER e JUANG, 1993). No entanto, os métodos de análise espectral são, provavelmente, os mais importantes.

Entre os métodos de análise espectral, encontram-se diversas técnicas, como, banco de filtros, banco de filtros digitais, coeficientes cepstrais e coeficientes de predição linear (PICONE, 1993). Tendo os dois últimos, grande destaque em aplicações em reconhecimento de fala.

Para melhor entender o processo de extração de característica, é importante ter em mente algumas considerações. Segundo (RABINER e JUANG, 1993), o sinal de fala é um sinal que varia lentamente no tempo, no sentido de que, quando examinado em um período suficientemente pequeno (entre 5 ms e 100 ms), suas características são relativamente estacionárias.

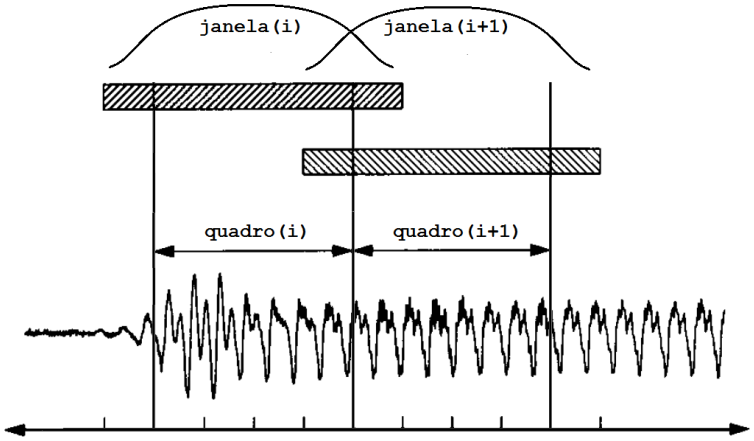
Dessa forma, pode-se recorrer à segmentação do sinal e, nesse caso, considerá-lo quase estacionário por partes. Essa segmentação pode ser vista como um processamento quadro-a-quadro, que consiste na segmentação do sinal a ser processado em pequenos conjuntos de amostras, seguido do janelamento destes

segmentos. De acordo com (PICONE, 1993), a duração dos segmentos está diretamente relacionada com a dinâmica do trato vocal. Para sinais de fala são usadas janelas de duração entre 10 ms e 40 ms.

A função do processo de janelamento é ponderar as amostras do sinal, valorizando as amostras do centro do segmento. Ainda segundo (PICONE, 1993), o janelamento, associado à sobreposição (entre os segmentos) tem a importante função de suavizar as variações das estimações paramétricas do sinal.

A Figura 13 ilustra os processos de janelamento e sobreposição utilizados na etapa de extração de características.

Figura 13 - Processo de segmentação: janelamento e sobreposição do sinal



Fonte: adaptado de (PICONE, 1993)

Deste modo, a extração de características do sinal é feita quadro-a-quadro, revelando assim, características que permitem modelar cada um dos segmentos do sinal. O vetor de característica é, na maior parte dos sistemas de reconhecimento de fala, resultado da concatenação de vetores de características menores, formados pelos parâmetros de cada um dos segmentos.

A seguir, são apresentados os princípios fundamentais dos principais métodos de extração de características para representação do sinal de fala. A Seção 4.1 trata, de forma mais generalista, os principais métodos existentes (embora existam muitos outros), enquanto a Seção 4.2 concentra-se em descrever os coeficientes mel-cepstrais, característica escolhida para representação dos sinais de fala neste trabalho. É comum encontrar trabalhos que fazem a utilização de combinações destes métodos.

4.1 PRINCIPAIS REPRESENTAÇÕES PARAMÉTRICAS DO SINAL

4.1.1 Energia e amplitude média em um curto intervalo de tempo

A energia e a amplitude média em um curto intervalo de tempo fornecem uma representação que reflete as variações de amplitude do sinal de fala no domínio do tempo. Em particular, a amplitude de segmentos de sons não vozeados é geralmente muito menor que a amplitude de segmentos vozeados (RABINER e SCHAFER, 1978). Desta forma, estas características permitem, principalmente, diferenciar um som vozeado de um som não vozeado.

O cálculo da energia em um curto intervalo de tempo é feito através da Equação (15):

$$\mathrm{E}_n = \sum_{m=-N/2}^{N/2} \mathrm{x}^2(\mathrm{m}) \cdot \mathrm{h}(\mathrm{n}-\mathrm{m}) \qquad \textbf{(15)}$$

Onde $\mathrm{x}(\mathrm{m})$ representa o sinal e $\mathrm{h}(\mathrm{n})$, a resposta ao impulso de um filtro linear e N, o comprimento da resposta ao impulso. Ou seja, o cálculo da energia do sinal corresponde à filtragem do sinal ao quadrado. A resposta ao impulso do filtro linear, $\mathrm{h}(\mathrm{n})$, é dada pela Equação (16):

$$h(n) = \begin{cases} w^2(n), & 0 \leq \mathrm{n} \leq \mathrm{N}-1 \\ 0, & \text{caso contário} \end{cases} \qquad \textbf{(16)}$$

A expressão $w(n)$ corresponde a uma janela, que pode ser retangular, de Hamming, de Hanning, entre outras. A escolha da janela está diretamente relacionada com a resposta em frequência desejada para o filtro linear utilizado no cálculo da energia.

O comprimento da janela influencia diretamente no resultado do cálculo da energia. Deve-se escolher um comprimento suficientemente pequeno, para que se possam representar as variações rápidas na amplitude do sinal, no entanto, um comprimento muito pequeno, não permite encontrar uma função de energia suavizada (RABINER e SCHAFER, 1978).

O cálculo da energia apresenta um problema, que é a sensibilidade a sinais de grande amplitude, pois é calculado pelo quadrado do sinal. Para resolver este problema, usa-se então, a função da amplitude média, expressa na Equação (17):

$$\mathrm{E}_n = \sum_{m=-N/2}^{N/2} |\mathrm{x(n)}| \cdot w(n-m) \qquad \textbf{(17)}$$

A utilização da função da amplitude média permite, além disso, evidenciar os sinais de amplitude muito baixa.

4.1.2 Taxa de cruzamento por zero

A taxa de cruzamento por zero – discutida anteriormente neste trabalho – é expressa pela Equação (5). Este índice representa a quantidade de vezes que a amplitude do sinal de fala sofreu alteração de sinal (de amplitudes positivas para amplitudes negativas ou vice-versa).

Assim como o cálculo da energia e da amplitude média, a taxa de cruzamento por zero é convenientemente utilizada na caracterização de um som entre vocálico ou não vocálico. Os sons não vocálicos apresentam maior parte da energia concentrada em frequências mais alta, implicando em ZCR mais altas, enquanto os sons vocálicos possuem maior parte de energia distribuída entre as baixas frequências, implicando em ZCR menores.

4.1.3 Frequência fundamental ou *pitch*

Segundo (HOWARD, 1992), os termos frequência fundamental e *pitch* são frequentemente confundidos. A frequência é um parâmetro físico, enquanto, o *pitch* é um parâmetro perceptual. Embora estas características estejam correlacionadas, a relação entre elas é complexa. No entanto, é comum, na literatura, que ambos os termos refiram-se à frequência fundamental.

A frequência fundamental é definida como a frequência de vibração das cordas vocais durante os sons vocálicos (HESS, 1983). De acordo com (SUN, 2000), a frequência fundamental é uma característica importante na modelagem e muitas outras áreas de pesquisa da fala. Uma grande quantidade de algoritmos de

determinação de *pitch* foi proposta no passado, entretanto, continua sendo um dos problemas mais difíceis em análise de fala.

Os principais algoritmos para determinação de *pitch* são brevemente discutidos em (PICONE, 1993). O autor cita ainda referências que apresentam os detalhes destes algoritmos.

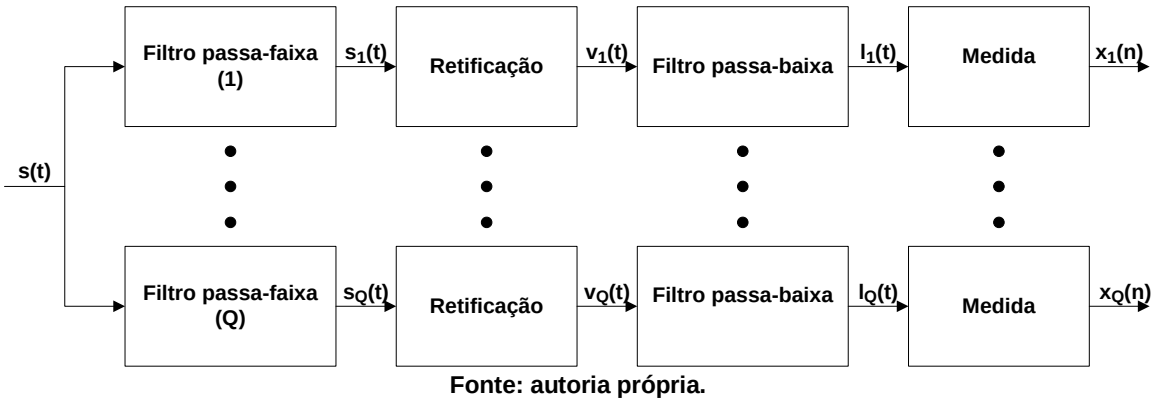
4.1.4 Banco de filtros

A extração de características baseada em banco de filtros foi uma das primeiras técnicas a ser utilizadas em reconhecimento automático de fala. Fortemente explorados até os anos 70, os bancos de filtros, segundo (PICONE, 1993), podem ser considerados como um modelo básico dos primeiros estágios da percepção da fala pelo sistema auditivo.

Os bancos de filtros podem ser implementados analógica e digitalmente, embora as primeiras aplicações tenham sido exclusivamente analógicas.

A implementação do banco de filtros analógicos, pode ser feita com uma série de filtros passa faixa, em paralelo. Cada filtro é ajustado para uma frequência central e banda passante específica. Após a filtragem o sinal é retificado (por um retificador de meia onda ou de onda completa) e, na sequência, novamente filtrado, por um filtro passa baixa, para que se possa medir a componente contínua – ou média – correspondente àquela banda de frequências. Como resultado do banco de filtros, podem-se tomar amostras das componentes médias para cada um dos *Q* filtros. A Figura 14 apresenta um diagrama de blocos da implementação analógica de banco de filtros.

Figura 14 - Diagrama básico da implementação analógica do banco de filtros.



Fonte: autoria própria.

A implementação digital do banco de filtros pode ser feita através da Transformada Rápida de Fourier ou através de filtros com resposta ao impulso de duração finita ou com resposta ao impulso de duração infinita (IIR). Em geral, faz-se o cálculo da energia em um curto intervalo de tempo para saída de cada um dos filtros.

Como saída, para ambas as implementações, têm-se um vetor formado por pares de frequência e energia ou amplitude média.

Além disso, os filtros (passa-faixa, na entrada do banco de filtros) podem ser distribuídos de maneira uniforme ou não, no domínio da frequência. A distribuição não uniforme é usada na tentativa de aproximar-se da escala da percepção das frequências pelo sistema auditivo. Em geral, são utilizadas as escalas *Bark* ou *mel* (PICONE, 1993).

As Equações (18) e (19) apresentam a expressão utilizada na conversão da frequência (em escala linear, dada em hertz) para as escalas *Bark* e *mel*, respectivamente.

$$\mathrm{Bark} = 13 \cdot \tan^{-1}\left(\frac{0.76 \cdot f}{1000}\right) + 3.5 \cdot \tan^{-1}\left(\frac{f^2}{7500^2}\right) \quad \textbf{(18)}$$

$$\mathrm{mel} = 2595 \cdot \log\left(1 + \frac{f}{700}\right) \quad \textbf{(19)}$$

A largura de banda crítica dos filtros pode ser calculada pela Equação (20), (PICONE, 1993).

$$\mathrm{BW}_{critico} = 25 + 75 \cdot \left[1 + 1.4 \cdot \left(\frac{f}{1000}\right)^2\right]^{0.69} \quad \textbf{(20)}$$

Onde f, neste caso, representa a frequência central. Uma tabela com valores padrão para a frequência central e a largura de banda dos filtros utilizados nos bancos de filtros foi definida em (ZWICKER e TERHARDT, 1980).

4.1.5 Coeficientes de predição linear

De acordo com (RABINER e SCHAFER, 1978), a codificação linear preditiva é uma das técnicas mais poderosas para análise de sinais de fala. Este método se

tornou o método predominante para estimação de parâmetros básicos da fala, como por exemplo, frequência fundamental, espectro, e forma do trato vocal, e para transmissão e armazenamento de sinais de fala em baixa taxa de bits. Ainda segundo o autor, a importância deste método está relacionada com a habilidade de fornecer parâmetros do sinal de fala extremamente precisos e com a velocidade com que os cálculos são realizados computacionalmente.

A ideia fundamental da codificação preditiva, é que cada amostra do sinal da fala pode ser estimada através de uma combinação linear das amostras passadas. Os coeficientes de predição linear – valores que ponderam as amostras passadas – são calculados pela minimização do erro entre a amostra real e a amostra estimada.

Ainda segundo (RABINER e SCHAFER, 1978), a filosofia da predição linear está intimamente ligada ao modelo de produção da fala (apresentado na Seção 2.4), que mostra que a fala pode ser descrita como uma saída de um sistema linear e variante no tempo, excitado por uma sequência de pulsos quase periódicos ou por ruído aleatório. A predição linear fornece uma estimativa precisa e confiável dos parâmetros que caracterizam este sistema.

A partir do modelo de produção de fala (vide Eq. (1)), podemos considerar este sistema linear como um filtro digital com função de transferência expressa pela Equação (21).

$$\mathrm{H}(z) = \frac{\mathrm{S}(z)}{\mathrm{U}(z)} = \frac{\mathrm{G}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad \textbf{(21)}$$

A amostra de fala, $\mathrm{s(n)}$, pode ser relacionada com o sinal de excitação, $u(n)$, através da equação de diferença, expressa na Equação (22). O índice p representa a ordem do filtro, e deve ser alto o suficiente para que o modelo forneça uma boa representação para a maioria dos sons.

$$\mathrm{s(n)} = \mathrm{G} \cdot \mathrm{u(n)} + \sum_{k=1}^{p} a_k \cdot \mathrm{s(n-k)} \qquad \textbf{(22)}$$

Como já foi apresentando, a técnica de predição linear consiste em estimar, através de uma combinação linear das amostras passadas do sinal, o valor da amostra atual. Deste modo, o sinal predito pode ser expresso pela Equação (23).

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k \cdot s(n-k) \quad \textbf{(23)}$$

Onde os coeficientes α_k representam os coeficientes de predição. O erro de predição é expresso pela Equação (24), e pode ainda ser considerado como a saída de um sistema linear com função de transferência $A(z)$, mostrada na Equação (25).

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k \cdot s(n-k) \quad \textbf{(24)}$$

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k \cdot z^{-k} \quad \textbf{(25)}$$

Comparando as Equações (22) e (23), é possível perceber que, se o sinal obedecer ao seu modelo de produção e se $\alpha_k = a_k$, então o erro de predição será $e(n) = G \cdot u(n)$. Dessa forma, o filtro de erro de predição, $A(z)$, será um filtro inverso do sistema, $H(z)$, isto é:

$$H(z) = \frac{G}{A(z)} \quad \textbf{(26)}$$

O problema básico da análise linear preditiva é encontrar o conjunto de coeficientes de predição que melhor aproximam o sinal estimado do sinal real. A busca deste conjunto de coeficientes é feita através da minimização do erro médio quadrático entre estes dois sinais.

De acordo com (RABINER e SCHAFER, 1978), existem diversos métodos para obtenção dos coeficientes de predição linear. O detalhamento destes métodos poderá ser encontrado em seu trabalho, entretanto, não serão abordados aqui.

4.2 COEFICIENTES MEL-CEPSTRAIS

Apesar de os coeficientes de predição linear representarem com precisão e confiabilidade as características do sinal de fala, diversos trabalhos mostram que a utilização de coeficientes cepstrais e mel-cepstrais apresentam melhores resultados nos sistemas de reconhecimentos automático de fala. Entre estes trabalhos pode-se destacar (AIDA-ZADE, ARDIL e RUSTAMOV, 2006) e (RUNSTEIN e VIOLARO, 1996).

Para definir os coeficientes mel-cepstrais (MFCC – *Mel-Frequency Cepstral Coefficients*), é interessante apresentar as ideias básicas da Transformada Discreta de Fourier, da representação através de coeficientes cepstrais, da escala Mel e da Transformada Discreta do Cosseno.

A Transformada Discreta de Fourier (DFT – *Discrete Fourier Transform*) consiste na Transformada de Fourier para sinais discretos e periódicos. Usualmente, é calculada através do algoritmo da FFT, apresentado inicialmente por (COOLEY e TUKEY, 1965). Maiores detalhes sobre estas transformadas podem ser encontrados em (LATHI, 2004) e (PROAKIS e MANOLAKIS, 1996).

A ideia da utilização da representação cepstral está relacionada com a natureza do sinal de fala. Como já foi apresentada, a produção do sinal de fala pode ser considerada um processo de filtragem, entre um sinal de excitação e um filtro linear variante no tempo, que representa o próprio trato vocal. A separação destas duas componentes – sinal de excitação e filtro – permite obter um modelo do trato vocal. Sabendo que a filtragem consiste na operação de convolução, tem-se:

$$s(n) = u(n) \otimes v(n) \tag{27}$$

Onde $u(n)$ denota a fonte de excitação e $v(n)$ a reposta ao impulso do trato vocal.

Conhecendo as propriedades da convolução, podemos representar a Equação (27) no domínio da frequência:

$$S(f) = U(f) \cdot V(f) \tag{28}$$

Aplicando o logaritmo complexo em ambos os lados, resulta em:

$$\log\big(S(f)\big) = \log(U(f) \cdot V(f))$$
$$\log\big(S(f)\big) = \log\big(U(f)\big) + \log(V(f)) \quad \textbf{(29)}$$

Desta forma, é possível separar o sinal de excitação e o modelo do trato vocal. Os coeficientes cepstrais são calculados por meio da transformada inversa de Fourier do logaritmo do espectro:

$$c(n) = \frac{1}{N_s} \cdot \sum_{k=0}^{N_s-1} \log|S(k)| \cdot e^{\left(\frac{2\pi}{N_s}\right) \cdot k \cdot n} \quad \textbf{(30)}$$
$$0 \leq n \leq N_s - 1$$

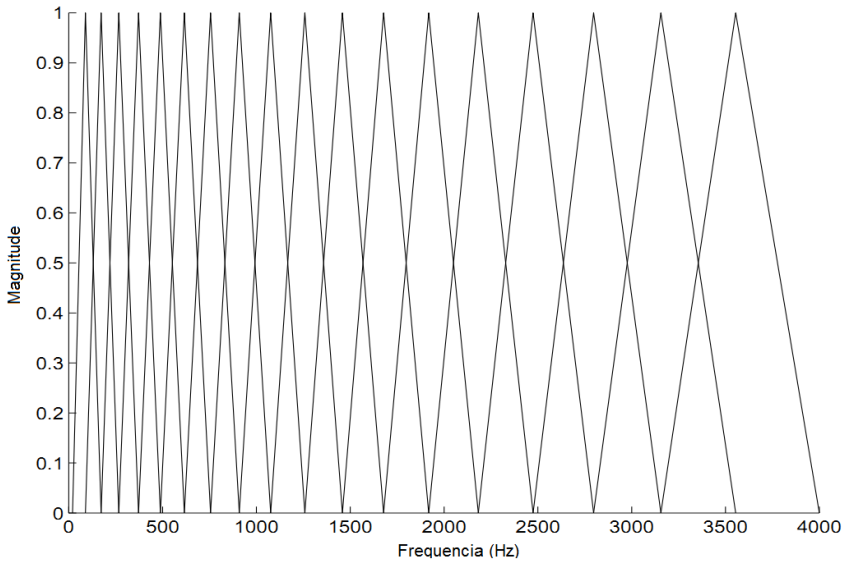
Os termos de baixa ordem dos coeficientes cepstrais estão relacionados com a forma do trato vocal, enquanto os temos de ordens mais elevadas, com o sinal de excitação. Na prática, para reconhecimento de fala, apenas os termos de baixa ordem *(n<20)* são utilizado, (PICONE, 1993). Ainda segundo o autor, o termo *c(0)* representa o valor médio do espectro, ou o valor RMS do sinal, e não trás informação útil para o reconhecimento de fala.

A escala *mel* – brevemente discutida na seção 4.1.4 – consiste em uma escala de frequências que tenta se aproximar da escala da percepção das frequências pelo sistema auditivo, mapeando, desta forma, o *pitch* do sinal, e não sua frequência. A escala *mel* foi definida experimentalmente e apresentada, pela primeira vez, por (STEVENS, VOLKMAN e NEWMAN, 1937).

A representação dos coeficientes cepstrais através da escala *mel* permite encontrar os coeficientes mel-cepstrais. O mapeamento dos coeficientes cepstrais através da escala *mel* é feito através de um banco de filtros mel.

O banco de filtros mel consiste em filtros triangulares com frequências de corte determinadas pela frequência central dos dois filtros adjacentes. Os filtros possuem frequências centrais espaçadas linearmente e largura de banda espaçadas na escala *mel* (COMBRINCK e BOTHA, 1996). A distribuição do banco de filtros em escala mel é mostrada na Figura 15.

Figura 15 - Banco de filtros triangulares em escala mel.



Fonte: adapatado de (COMBRINCK e BOTHA, 1996)

A Transformada Discreta do Cosseno (DCT – *Discrete Cosine Transform*) foi introduzida, inicialmente, por (AHMED, NATARAJAN e RAO, 1974) e consiste em uma transformação que projeta no domínio da frequência, um sinal descrito no domínio do tempo, assim como a DFT. Entretanto, com o uso da DCT é possível agrupar grande parte da energia do sinal nos coeficientes de baixa ordem, permitindo que o sinal – no domínio do tempo - possa ser reconstruído a partir de poucos coeficientes da DCT, o que nem sempre é possível com o uso da DFT. Além disso, o cálculo da DCT resulta apenas em coeficientes reais, enquanto o cálculo da DFT resulta em valores complexos (BLINN, 1993). A transformada discreta do cosseno é frequentemente aplicada à compactação de imagens.

Tendo conhecido os conceitos intermediários, pode-se proceder com o cálculo dos MFCC. Em (RUNSTEIN, 1998) são listados os principais algoritmos utilizados no cálculo dos coeficientes mel-cepstrais. Os principais passos destes algoritmos são listados a seguir:

- Algoritmo de Davis e Mermelstein, (DAVIS e MERMELSTEIN, 1980):
  - FFT das amostras do segmento de fala;
  - Cálculo do quadrado do módulo da FFT;
  - Filtragem do resultado acima por um banco de filtros na escala *mel*;

- Cálculo do logaritmo da energia na saída dos filtros;
- Cálculo da DCT, resultando nos MFCC.

- Algoritmo de Deller, Proakis e Hansen, (DELLER, HANSEN e PROAKIS, 1993):
  - FFT das amostras do segmento de fala;
  - Cálculo do logaritmo do quadrado da FFT;
  - Filtragem do resultado acima por um banco de filtros na escala *mel*;
  - Cálculo da DFT, resultando nos MFCC.

5 RECONHECIMENTO DE PADRÕES

A etapa final do sistema de reconhecimento de fala consiste no reconhecimento/classificação dos padrões, a partir da apresentação do vetor de características.

O reconhecimento de padrões (RP) pode ser definido de diversas formas. Em (LIU, SUN e WANG, 2006) são listadas dezenas de autores e suas definições para o problema. Na definição mais recente, o reconhecimento de padrão é apresentado como uma ciência, cujo objetivo é classificar um objeto (sinal, imagem, etc.), baseado em suas características, em um padrão ou classe, dentre um conjunto de padrões (THEODORIDIS e KOUTROUMBAS, 2008)

A tarefa de reconhecimento de padrões pode ser realizada por um conjunto de técnicas matemáticas, estatísticas, heurísticas e indutivas. Além disso, segundo (SÁ, 2001), as principais abordagens dos problemas de RP são: abordagem estatística, abordagem sintática, abordagem neural e abordagem difusa.

Neste trabalho utilizar-se-á a abordagem neural, em especial, baseada em redes neurais artificiais. Para maiores informações sobre o problema de RP e sobre as outras abordagens para o tema, recomenda-se consultar (THEODORIDIS e KOUTROUMBAS, 2008) e (SÁ, 2001).

A seguir serão apresentados os principais conceitos relacionados às redes neurais artificiais e sua aplicação em reconhecimento automático de fala.

5.1 REDES NEURAIS ARTIFICIAIS

As redes neurais artificiais, RNA, de acordo com (SILVA, SPATTI e FLAUZINO, 2010) são modelos computacionais inspirados no sistema nervoso dos seres vivos, possuem a capacidade de aquisição e manutenção do conhecimento e podem ser definidas por um conjunto de unidades de processamento, chamados de neurônios artificiais, interligados por um grande número de interconexões, representados por meio de pesos sinápticos.

Segundo (HAYKIN, 2001, p. 28), a rede neural artificial é definida como:

> "... um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso. Ela se

> assemelha ao cérebro em dois aspectos: (1) o conhecimento adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem; (2) forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido."

As principais características apresentadas pelas redes neurais são:

- Aprendizado por experiência – o ajuste dos parâmetros internos da rede é feito através da apresentação de exemplos (experiências anterior) relacionados com o comportamento do processo;
- Capacidade de aprendizado – a partir de um algoritmo de aprendizado, a rede neural é capaz de extrair conhecimento sobre o processo onde é aplicada;
- Habilidade de generalização – a partir do conhecimento extraído sobre o processo, a rede é capaz de generalizar este conhecimento e assim, encontrar soluções mesmo para dados de entrada que não tenham sido usados no seu treinamento;
- Adaptabilidade – a rede neural pode ser utilizada em processos não-estacionários, através de um treinamento em tempo real, permitindo assim que a rede continue apresentando um bom desempenho após variações do processo;
- Não-linearidade – a utilização de unidades de processamento (neurônios) não lineares fornece à rede neural esta característica, que permite que ela seja utilizada, também, em processos não lineares;
- Armazenamento distribuído – o conhecimento adquirido à respeito do processo é armazenado, de forma distribuída, nas conexões entre os neurônios;
- Tolerância a falhas – em função da sua grande interconexão, a rede neural é capaz de manter-se operando satisfatoriamente mesmo que sua estrutura sensivelmente corrompida.

Em função de todas estas características, as redes neurais podem ser aplicadas à problemas de aproximação de função, previsão de séries temporais, controle de processos, agrupamento de dados, otimização, reconhecimento de padrões, entre outros

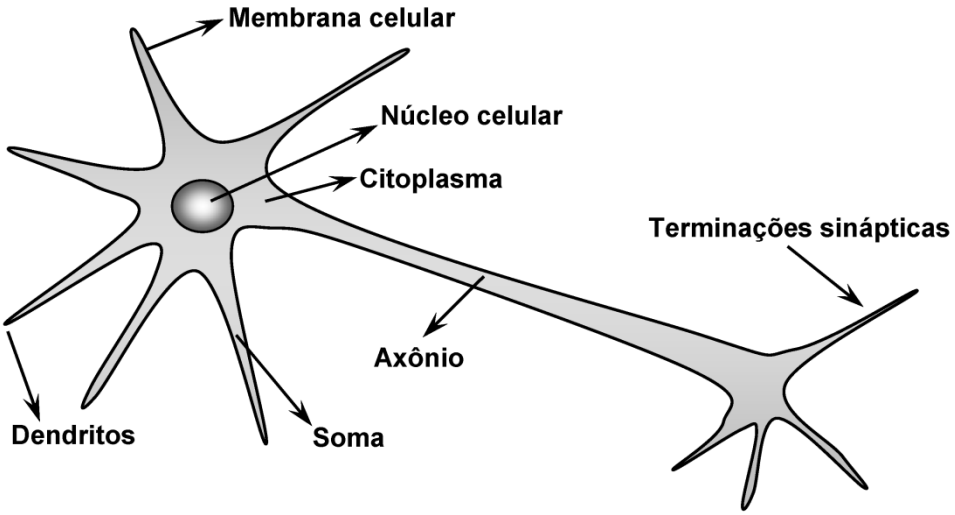
Entre estas características, as que as tornam a RNA interessante para aplicações em reconhecimento de fala são, principalmente, o aprendizado por experiência, a capacidade de aprendizado e a habilidade de generalização.

As redes neurais podem apresentar diversas arquiteturas e topologias, além de diversos algoritmos de treinamento. A seguir serão apresentadas as informações mais relevantes para o entendimento das RNA.

5.1.1 Neurônios

O neurônio artificial constitui a unidade básica de processamento de uma rede neural. Seu modelo é uma aproximação do neurônio biológico, a célula elementar do sistema nervoso cerebral. O neurônio biológico é constituído de três partes principais: os dendritos, o corpo celular e o axônio. A Figura 16 ilustra um destes neurônios.

Figura 16 - Neurônio biológico.



Fonte: (SILVA, SPATTI e FLAUZINO, 2010)

Os dendritos têm como principal função receber a informações do meio externo através de neurônios sensitivos, e ou de outros neurônios os quais esteja conectado.

O corpo celular concentra outros elementos celulares (núcleo, mitocôndria, lisossomo, etc.) e tem por função processar o sinal recebido pelos dendritos e gerar um potencial de ativação para disparar impulsos elétricos a outros neurônios.

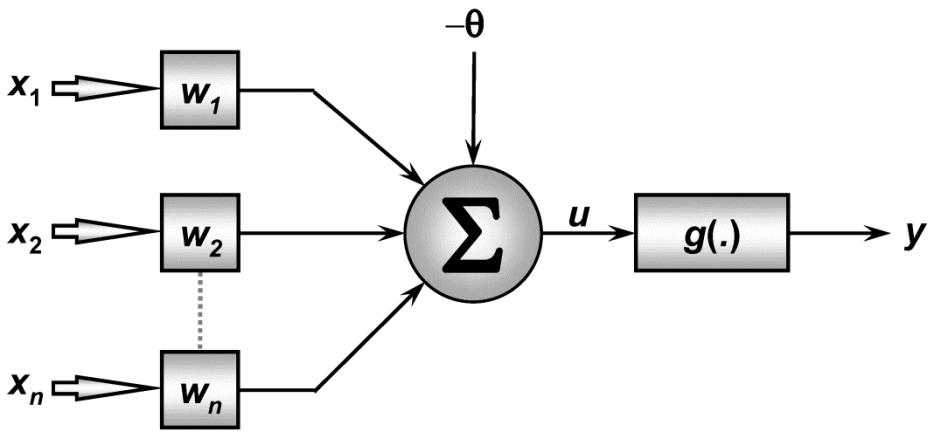
O axônio consiste em um prolongamento do neurônio, que agrega ramificações, conhecidas como terminações sinápticas, e é responsável por transmitir os impulsos elétricos produzidos pelo corpo celular à outros neurônios intermediários ou à neurônios ligados ao tecido muscular (neurônios efetuadores).

A comunicação entre os neurônios é feita através das conexões sinápticas. A sinapse é realizada, nos neurônios biológicos, através de um processo químico, que libera substâncias neurotransmissoras.

A partir do conhecimento do neurônio biológico, (MCCULLOCH e PITTS, 1943) apresentaram um modelo de neurônio artificial, considerado o mais simples e mais utilizado ainda hoje.

Os neurônios artificiais utilizado nas redes neurais são, tipicamente, não lineares, fornecem geralmente saídas contínuas e realizam funções simples, como agregar os sinais existentes em suas entradas de acordo com sua função operacional e produzir uma resposta, levando em consideração sua função de ativação (SILVA, SPATTI e FLAUZINO, 2010). A Figura 17 apresenta uma ilustração do modelo do neurônio artificial.

Figura 17 - Neurônio artificial.



Fonte: (SILVA, SPATTI e FLAUZINO, 2010).

Considerando a Figura 17, verifica-se a existência de sete elementos básicos do neurônio artificial (SILVA, SPATTI e FLAUZINO, 2010):

a) Sinais de entrada $\{x_1, x_2,..., x_n,\}$: são os sinal advindos do ambiente ou do meio externo, contendo as informações a serem processadas pela rede neural;

b) Pesos sinápticos {w_1, w_2,..., w_n,}: são os valores que ponderarão os sinais de entrada, permitindo quantificar a relevância do sinal para o processamento do neurônio;
c) Combinador linear {Σ}: soma os sinais de entrada do neurônio, já ponderado pelos pesos sinápticos para produzir um potencial de ativação;
d) Limiar de ativação {θ}: especifica um limiar para que o resultado produzido pelo combinador linear possa gerar um disparo em direção aos neurônios seguintes;
e) Potencial de ativação {u}: é o resultado da diferença entre o valor produzido pelo combinador linear e o limiar de ativação;
f) Função de ativação {g}: função que restringe a amplitude do sinal de saída de um neurônio;
g) Sinal de saída {y}: consiste no resultado do valor do potencial de ativação aplicado à função de ativação do neurônio. Este sinal pode ser levado diretamente à saída da rede, ou então, ser utilizado como entrada de outros neurônios.

Dessa forma, o sinal de saída do neurônio pode ser calculado através das Equações (31) e (32).

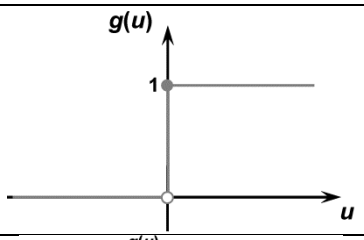
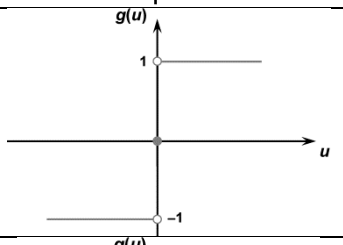
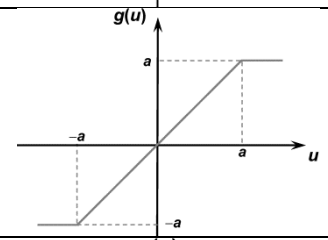
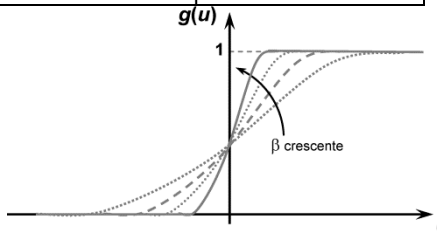
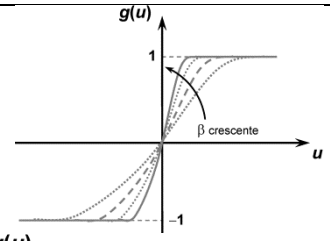
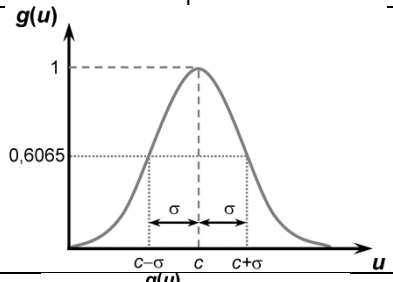
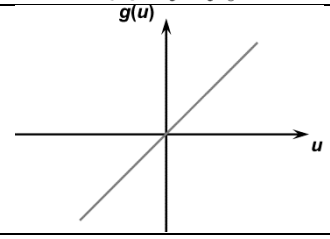
$$u = \sum_{i=1}^{n} (w_i \cdot x_i) - \theta \quad \textbf{(31)}$$

$$y = g(u) \quad \textbf{(32)}$$

As principais funções de ativação utilizadas nos neurônios das redes neurais são mostradas na Tabela 1.

Uma consideração importante a ser feita quanto aos valores de entrada da RNA é que a normalização destes valores permitindo um incremento de desempenho e eficiência computacional dos algoritmos de aprendizagem.

Tabela 1- Funções de ativação do neurônio artificial.

Funções de ativação		
Função degrau	$g(u) = \begin{cases} 1, & se\ u \geq 0 \\ 0, & se\ u < 0 \end{cases}$	
Função degrau bipolar	$g(u) = \begin{cases} 1, & se\ u > 0 \\ 0, & se\ u = 0 \\ -1, & se\ u < 0 \end{cases}$ ou $g(u) = \begin{cases} 1, & se\ u \geq 0 \\ -1, & se\ u < 0 \end{cases}$	
Função rampa simétrica	$g(u) = \begin{cases} a, & se\ u > a \\ u, & se\ -a \leq u \leq a \\ -a, & se\ u < a \end{cases}$	
Função logística	$g(u) = \frac{1}{1 + e^{-\beta u}}$	
Função tangente hiperbólica	$g(u) = \frac{1 - e^{-\beta u}}{1 + e^{-\beta u}}$	
Função gaussiana	$g(u) = e^{-\frac{(u-c)^2}{2\sigma^2}}$	
Função linear	$g(u) = u$	

Fonte: adaptado de (SILVA, SPATTI e FLAUZINO, 2010).

5.1.2 Arquitetura e treinamento de redes neurais artificiais

Uma rede neural pode ser descrita em termos de sua arquitetura e sua topologia. A arquitetura está relacionada com o arranjo dos neurônios em relação uns aos outros e com o direcionamento das ligações sinápticas. A topologia descreve as diferentes composições de uma arquitetura.

As redes neurais são, em sua maioria, formadas por três camadas: camada de entrada, camadas intermediárias e camada de saída.

A camada de entrada é aonde chegam os dados provenientes do meio externo. As camadas intermediárias são compostas por neurônios responsáveis pela extração do conhecimento sobre o processo. A maior parte do processamento é realizado nestas camadas. Por fim, a camada de saída é responsável por processar e apresentar os resultados provenientes do processamento realizado nas camadas intermediárias.

As arquiteturas das redes neurais estão relacionadas, principalmente, a dois parâmetros: a propagação dos sinais e a estrutura das camadas.

Quanto à estrutura das camadas, podem ser consideradas redes de camada simples (onde não há camada intermediária, apenas as camadas de entrada e saída), múltiplas camadas (com uma ou mais camadas intermediárias) e redes reticuladas, composta por neurônios dispostos espacialmente.

Quanto à propagação dos sinais, as redes podem ser classificadas entre *feedforward*, onde os sinais seguem num único sentido – da camada de entrada para a camada de saída – e recorrente, ou realimentada, onde os sinais da camada de saída podem servir de entrada em camadas anteriores – percorrendo o sentido reverso.

O treinamento da rede neural consiste no processo de aquisição do conhecimento, realizado através de um algoritmo específico e resulta na alteração dos pesos sinápticos das conexões entre neurônios. Esta etapa pode ser classificada de duas formas: treinamento supervisionado e treinamento não supervisionado.

No primeiro, a rede neural é treinada a partir de um conjunto de dados contendo informações da entrada e da respectiva saída esperada. Esse tipo de treinamento é considerado um caso típico de inferência indutiva. Uma variação do treinamento supervisionado é o treinamento com reforço, onde não se conhece os

valores desejados para as saídas, mas tem-se disponível uma informação quantitativa ou qualitativa da reposta, indicando apenas se o resultado é ou não satisfatório.

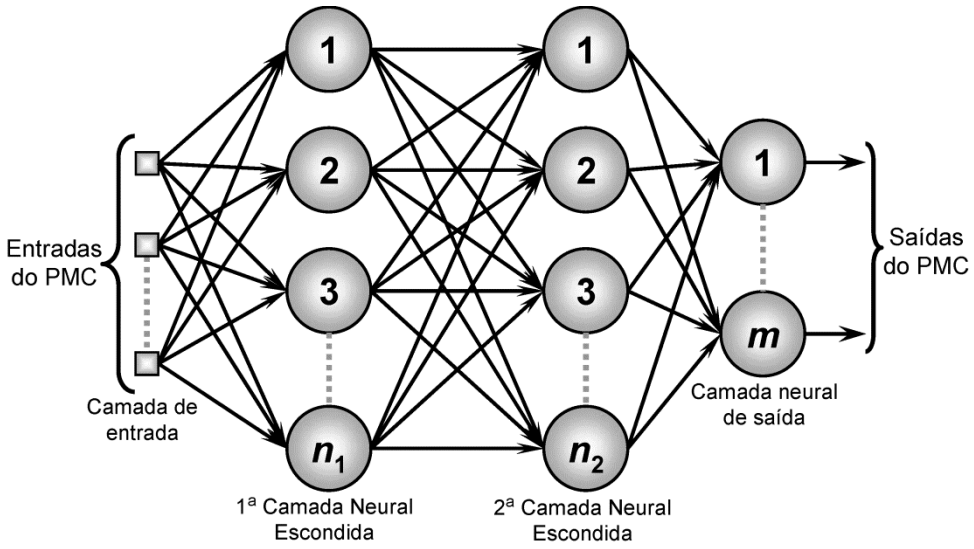
Já no treinamento não supervisionado, não se tem conhecimento prévio das saídas desejadas. Deste modo, a rede deve se auto organizar para encontrar similaridades, identificando subconjuntos entre o conjunto de dados apresentado.

5.1.3 Rede Perceptron Multicamandas

Existem diversas arquiteturas de redes neurais disponíveis na literatura científica, no entanto, a rede Perceptron Multicamadas – PMC, ou MLP, do inglês, *Multi-Layer Perceptron* – é uma das mais utilizadas e é altamente versátil, quanto à aplicabilidade. Por ser a arquitetura escolhida para utilização neste trabalho, as demais arquiteturas não serão aqui detalhadas.

As redes MLP são redes com arquiteturas *feedforward* e múltiplas camadas, cujo treinamento é realizado de forma supervisionada. A Figura 18 ilustra a arquitetura de uma rede perceptron multicamadas.

Figura 18 - Rede Perceptron Multicamadas.



Fonte: (SILVA, SPATTI e FLAUZINO, 2010)

De acordo com (SILVA, SPATTI e FLAUZINO, 2010), a configuração topológica da rede (número de camadas e número de neurônios nas camadas) está

relacionado com o problema as ser tratado, com a disposição espacial das amostras e com os valores iniciais atribuídos aos pesos sinápticos.

O treinamento da rede PMC usualmente é feito através do algoritmo *backpropagation* ou retro propagação do erro. A abordagem matemática deste algoritmo não será apresentada neste trabalho, mas pode ser encontrada com detalhes em (HAYKIN, 2001) e (SILVA, SPATTI e FLAUZINO, 2010).

O treinamento da rede Perceptron Multicamadas consiste basicamente num problema de minimização de erro. Este processo é realizado através da realização sucessiva de duas fases distintas. A primeira fase é a “propagação a frente”, onde são apresentados à rede, os sinais de entrada. Os sinais de entrada são processados pelos neurônios das camadas intermediarias e se propagam até a camada de saída, produzindo os resultados da rede neural. Os resultados obtidos nesta fase são comparados com os resultados esperados para tais valores de entrada. O erro – diferença entre os resultados e os valores esperados – é utilizado para avaliar o desempenho da rede e também serve de parâmetro para a segunda fase do algoritmo.

A segunda fase é chamada de propagação reversa, e consiste no ajuste dos pesos sinápticos a partir do valor de erro para cada uma das saídas. O ajuste dos pesos sinápticos é feito em relação à direção oposta ao gradiente da função de erro quadrático, e calculado através da soma do valor dos pesos atuais com um valor, conhecido como passo de atualização, proporcional à derivada do erro e a uma taxa, chamada taxa de aprendizagem. Entretanto, só é possível determinar o erro na camada de saída. O ajuste dos pesos das camadas intermediárias é feito através de uma estimativa do erro, fornecidas pela camada imediatamente à frente. Ou seja, a última camada fornece uma estimativa do erro à penúltima camada, que por sua vez, fornece uma estimativa à antepenúltima, e assim sucessivamente.

As execuções repetidas dessas duas fases fazem com que, após diversas atualizações, os pesos sinápticos sejam estabilizados, permitindo assim, que estes valores (dos pesos) representem o conhecimento adquirido pela rede neural. Após o treinamento, a rede neural pode ser colocada em operação. Nesta etapa os pesos sinápticos já não sofrem mais alterações.

Desde sua apresentação, diversas modificações e adaptações foram propostas ao algoritmo de aprendizagem *backpropagation*. Uma destas propostas é o algoritmo “*resilient propagation”*, ou RPROP (RIEDMILLER e BRAUN, 1993).

No algoritmo *backpropagation*, o ajuste dos pesos sinápticos é proporcional ao gradiente da função de erro, entretanto, em alguns casos tais derivadas parciais podem assumir valores muito pequenos – causando variações muito pequenas no ajuste dos pesos, o que reduz a velocidade de convergência – mesmo com os pesos longe dos valores ótimos. O propósito do algoritmo RPROP é eliminar este problema.

No algoritmo de treinamento *resilient propagation* o ajuste dos pesos não é influenciado pela magnitude das derivadas, apenas pelo seu sinal. Além disso, os passos de atualização são adaptativos, sendo incrementados quando o sinal do gradiente da função de erro com relação aos pesos mantém-se igual durante duas iterações sucessivas e decrementado quando há troca de sinal.

Neste trabalho optou-se por utilizar o algoritmo RPROP no treinamento das redes neurais, pois, tal algoritmo apresenta maior velocidade de convergência e robustez a escolha dos parâmetros iniciais (RIEDMILLER e BRAUN, 1993).

5.1.4 O reconhecimento de padrões com redes neurais artificiais

Um dos problemas para quais as redes PMC se mostram úteis são os problemas de reconhecimento/classificação de padrões.

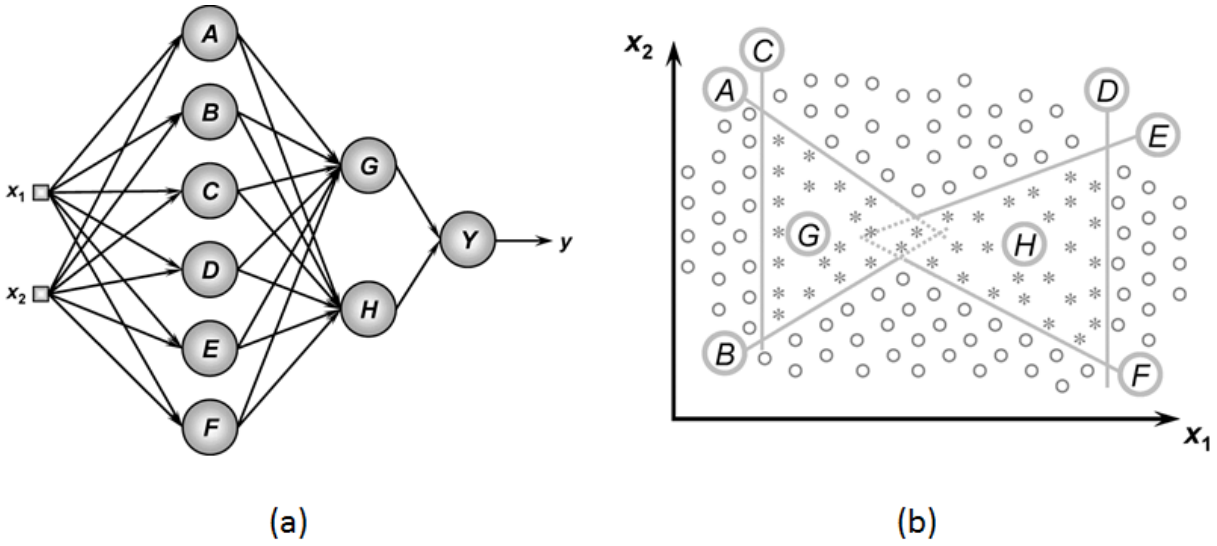
Um aspecto relevante nesta classe de problemas é que as saídas das redes neurais usadas em classificação de padrões são, tipicamente, binárias, indicando, por exemplo, se determinado padrão de entrada pertence à uma classe (nível lógico “1”) ou não pertence (nível lógico “0”).

Nos problemas de classificação de padrões, os neurônios da rede neural formam fronteiras de separabilidade. Tais fronteiras consistem em hiperplanos, no espaço vetorial que contém os sinais de entrada da rede. A Figura 19 ilustra um exemplo de problema de classificação de padrões, cuja entrada dos dados é bidimensional. Neste caso, a fronteira de separabilidade é definida por uma reta. Na Figura 19 (a) é mostrada a arquitetura da rede neural utilizada, enquanto na Figura 19 (b) são mostradas as fronteiras de separabilidade correspondentes a cada neurônio da rede.

Um dos métodos mais utilizados na representação das saídas das redes neurais aplicadas ao reconhecimento de padrão é conhecido como *one of c-classes* (SILVA, SPATTI e FLAUZINO, 2010). Nesse método, cada saída da rede está associada a uma dos padrões.

Em muitos casos é necessária uma etapa de pós-processamento dos sinais de saída de rede neural. Nesta etapa, por exemplo, as saídas podem assumir a forma binária, recebendo valor lógico “1” a saída que apresentar o maior valor, e nível lógico “0” todas as outras.

Figura 19 –Problema de classificação de padrões com RNA: fronteiras de separabilidade.



Fonte: adaptado de (SILVA, SPATTI e FLAUZINO, 2010)

6 EXPERIMENTOS

6.1 CONFIGURAÇÃO DOS EXPERIMENTOS

6.1.1 Base de dados

A base de dados para treinamento, teste e validação do sistema foi gravada a parir de um computador pessoal, usando uma placa de áudio *IDT High Definition Audio* e o microfone *Microsoft® LifeChat™ LX-2000*. Durante as gravações o microfone foi mantido estático, a uma distância entre 15cm e 20cm do usuário. As gravações foram feitas por meio do software *Audaticy 1.3®*, a uma frequência de amostragem de 11025 Hz, resolução de 16 bits e duração de um segundo, e salvas em arquivos no formato ".wav". Nenhum tratamento do sinal (e. g., filtragem, remoção do ruído) foi feita na etapa de gravação da base de dados.

O ambiente no qual foram realizadas as gravações foi uma sala com ruído ambiente, gerado principalmente por aparelho de ar-condicionado e computadores.

Utilizou-se um total de 30 usuários, sendo estes divididos em 20 usuários do sexo masculino e 10, do sexo feminino. Os usuários possuíam naturalidade de diversas cidades da região sul do Brasil, apresentando, assim, diferentes sotaques.

O conjunto de dados com as gravações de áudio foi formado pelas pronúncias das palavras: zero, um, dois, três, quatro, cinco, seis, sete, oito, nove, esquerda, direita, frente, atrás, sobe, desce, liga, desliga, acende, apaga, abre, fecha, casa, carro, luz e porta. Foram gravadas cinco pronúncias de cada palavra para cada um dos usuários, totalizando 3900 arquivos de áudio.

6.1.2 Pré-processamento e Extração de Características

Os algoritmos de pré-processamento e extração de características foram todos desenvolvidos através do software MATLAB®.

Na etapa de pré-processamento, os sinais de áudio – da base de dados construída – foram filtrados por um filtro passa-faixa, com uma banda de passagem entre 100 Hz e 4.000 Hz de ordem 200, e, em seguida, filtrados pelo filtro de pré-ênfase. No filtro de pré-ênfase foi utilizado $a_{pre} = 0.95$. Após estes procedimentos, o sinal foi normalizado, para então ser extraído apenas o intervalo correspondente à pronúncia da palavra.

Para detecção de palavras foi utilizado o algoritmo proposto na Seção 3.3.3. Foram utilizados segmentos de aproximadamente 3.6 ms – correspondente à 40 amostras do sinal, com sobreposição de 50% – no cálculo da energia. O limiar de detecção foi ajustado em um valor correspondente a quatro vezes a mediana da energia do sinal.

Na etapa de extração de caraterísticas foram calculados os coeficientes MFCC. O sinal de áudio, resultante da extração da palavra, foi dividido em 20 segmentos, com duração variável (conforme a duração do sinal) e com sobreposição de 33,3%.

O cálculo dos coeficientes foi feito através das funções disponíveis no Toolbox “VOICEBOX: Speech Processing Toolbox for MATLAB”, (BROOKES, 2005). O algoritmo utilizado no cálculo dos coeficientes é baseado no algoritmo de Davis e Mermelstein.

Os segmentos do sinal foram ponderados por uma janela de Hamming, que, segundo (PICONE, 1993) é a mais utilizada em reconhecimento de fala. O número de filtros na composição do banco de filtros *mel* – com forma triangular – e o número de coeficientes MFCC calculados foram ajustados de acordo com os experimentos. Não foi considerando o coeficiente *c(0)*. Os conjuntos de coeficientes foram concatenados, formando um único vetor de características. A fim de aperfeiçoar o desempenho da RNA, o vetor características é sempre normalizado.

A utilização de redes neurais para classificação de padrões impõe uma limitação ao sistema de reconhecimento de fala. Como a topologia da rede é, em geral, uma topologia fixa, ou seja, os números de entradas, de camadas e de neurônios, são escolhidos no início do problema e são mantidos fixos, são necessários alguns procedimentos para contornar esta limitação.

Uma opção é encontrar a palavra de maior duração, definir o número de segmentos do qual serão extraídas as características e fixar o número de segmentos e sua duração. Desta forma, para os sinais com duração menor, o vetor de características é inicialmente formado pelas informações da palavra nas posições e preenchido com zeros até a posição que corresponde a duração da maior palavra.

Outra opção poderia ser a reamostragem do sinal (através do processo de interpolação) com taxas de amostragem variável, a fim de que todos os sinais possuam o mesmo número de amostras.

A opção escolhida para resolver tal problema neste trabalho foi a utilização de um número fixo de segmentos, com duração variável. Dessa forma, a duração de cada segmento é ajustada dinamicamente, com relação à duração total da palavra, de modo que toda a palavra possa ser descrita pelo número de segmentos escolhidos.

Como comentado anteriormente, escolheu-se utilizar 20 segmentos para representar o sinal de fala. A utilização de 20 segmentos garante ao algoritmo de extração de característica que os quadros tenham duração entre 10 e 40 ms para praticamente todas as palavras, assegurando assim a quase estacionariedade dos segmentos.

6.1.3 Classificação de Padrões

As redes neurais utilizadas no processo de classificação de padrões foram redes MLP, treinadas a partir do algoritmo *resilient propagation*.

O conjunto de dados foi dividido randomicamente, sendo que, para o treinamento foram utilizados 80% do conjunto e os 20% restantes foram utilizados para teste. Como critério de parada para o algoritmo de treinamento foram utilizados número de épocas igual a 500 ou módulo do gradiente menor que $1x10^{-10}$.

A rede projetada possui apenas uma camada intermediária e o número de neurônios na camada intermediária foi selecionado de acordo com o experimento. O número de neurônios na camada de saída corresponde sempre ao número de classes a serem classificadas.

As redes neurais foram projetadas criadas e configuradas a partir da ferramenta Neural Network Toolbox (DEMUTH, BEALE e HAGAN, 2010).

Em todos os experimentos foram realizados 30 treinamentos. Ao final, foram calculadas a taxa de acerto média e o seu desvio padrão.

6.2 EXPERIMENTO 1

Neste experimento foi feita uma análise comparativa entre o algoritmo de detecção de palavras proposto no trabalho e a versão original, proposta por (BRESOLIN, 2008).

Os conjuntos de dados foram formados por 12 coeficientes MFCC, obtidos a partir de um banco de filtros *mel* com 13 filtros triangulares, para cada segmento. Foram criados quatro conjuntos de dados:

- Todas as palavras usando o algoritmo original;
- Todas as palavras usando o algoritmo proposto;
- Apenas dígitos usando o algoritmo original;
- Apenas dígitos usando o algoritmo proposto;

Todos estes conjuntos de dados foram classificados a partir de uma RNA com 50 neurônios na camada intermediária.

6.3 EXPERIMENTO 2

O objetivo deste experimento é avaliar a influência do número de coeficientes MFCC na taxa de acerto do sistema. Foram utilizados os seguintes conjuntos de dados:

- Todas as palavras e 12 coeficientes MFCC por segmento;
- Apenas dígitos e 12 coeficientes MFCC por segmento;
- Todas as palavras e 14 coeficientes MFCC por segmento;
- Apenas dígitos e 14 coeficientes MFCC por segmento;
- Todas as palavras e 16 coeficientes MFCC por segmento;
- Apenas dígitos e 16 coeficientes MFCC por segmento;
- Todas as palavras e 18 coeficientes MFCC por segmento;
- Apenas dígitos e 18 coeficientes MFCC por segmento;
- Todas as palavras e 20 coeficientes MFCC por segmento;
- Apenas dígitos e 20 coeficientes MFCC por segmento;

Em todos estes conjuntos foram utilizados o algoritmo de detecção de palavras proposto e uma RNA com 50 neurônios na camada intermediária.

6.4 EXPERIMENTO 3

Neste experimento foi avaliada a influência do número de neurônios da camada intermediária na taxa de acerto do sistema. Foram utilizados os seguintes conjuntos de dados:

- Todas as palavras e 20 neurônios na camada intermediária;
- Apenas dígitos e 20 neurônios na camada intermediária;
- Todas as palavras e 50 neurônios na camada intermediária;
- Apenas dígitos e 50 neurônios na camada intermediária;
- Todas as palavras e 80 neurônios na camada intermediária;
- Apenas dígitos e 80 neurônios na camada intermediária;

Em todos estes conjuntos foram utilizados o algoritmo de detecção de palavras proposto e extraídos 12 coeficientes MFCC.

7 RESULTADOS E DISCUSSÕES

7.1 EXPERIMENTO 1

No primeiro experimento foram comparados o algoritmo de detecção de palavras original, apresentado em (BRESOLIN, 2008) e sua modificação, proposta neste trabalho.

A Tabela 2 apresenta os resultados obtidos neste experimento. Na tabela são mostradas as taxas de acerto médias e seus respectivos desvios padrões para ambos os algoritmos utilizando como conjunto de dados todas as palavras ou apenas as palavras correspondente aos dígitos (zero, um,..., nove).

Tabela 2 - Resultados: Experimento 1

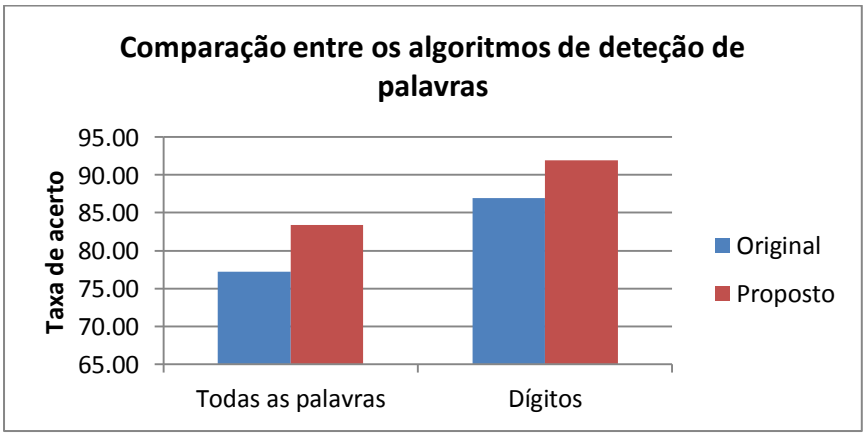
	Taxa de acerto			
	Todas as palavras		Dígitos	
Algoritmo	Média	Desvio Padrão	Média	Desvio Padrão
Original	77.17%	1.67%	86.98%	2.25%
Proposto	83.39%	3.22%	91.92%	2.64%

Fonte: autoria própria.

Através dos resultados obtidos é possível verificar que a modificação proposta neste trabalho para o algoritmo apresentado por (BRESOLIN, 2008) permitiu incrementar a taxa de acerto na tarefa de classificação dos padrões.

O resultado do experimento é ilustrado também através do Gráfico 1.

Gráfico 1 - Resultados do Experimento 1



Fonte: autoria própria.

Quando analisados os resultados obtidos para o conjunto completo de palavras, o algoritmo proposto proporcionou um aumento de 6.22% na taxa de acerto. Para o conjunto de palavras contendo apenas os dígitos, o incremento foi de 4.94%.

A superioridade do algoritmo apresentado deve-se ao fato de que este algoritmo concentra melhor a palavra no sinal extraído da gravação, enquanto no algoritmo original pode ocorrer extração de sinais de silêncio (entre o som da batida dos lábios e o início real da palavra, por exemplo) junto com o intervalo correspondente à palavra, como mostrou a Figura 12.

Outra característica que contribui para a menor taxa de acerto do algoritmo original é que os sons relativos aos elementos externos e/ou não estacionários, que prejudicam a detecção da palavra, não ocorrem igualmente em todas as palavras ou até mesmo nas diversas pronúncias da mesma palavra.

7.2 EXPERIMENTO 2

O segundo experimento teve por objetivo avaliar a influência do número de coeficientes MFCC extraídos de cada segmento do sinal de fala, utilizados para compor o vetor de características. Os resultados do experimento são mostrados na Tabela 3.

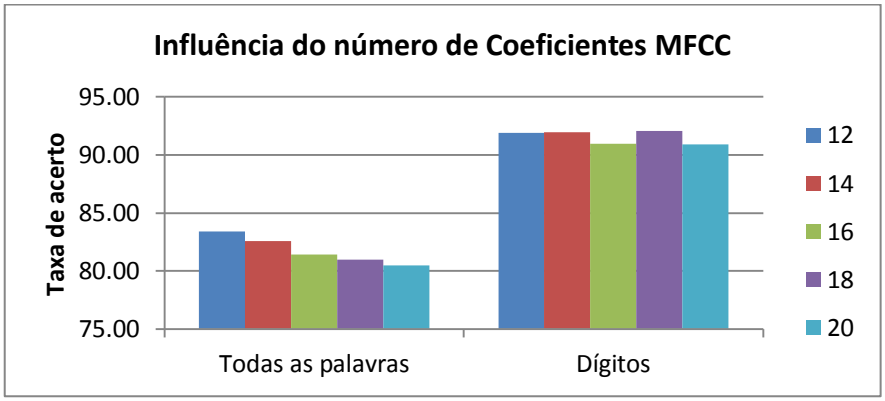
Tabela 3 - Resultados: Experimento 2

Número de coeficientes MFCC	Taxa de acerto			
	Todas as palavras		Dígitos	
	Média	Desvio Padrão	Média	Desvio Padrão
12	83.39%	3.22%	91.92%	2.64%
14	82.59%	2.55%	91.97%	1.83%
16	81.42%	2.00%	90.96%	3.43%
18	81.00%	2.43%	92.09%	2.93%
20	80.46%	2.43%	90.93%	2.55%

Fonte: autoria própria.

O Gráfico 2 apresenta uma ilustração dos resultados obtidos.

Gráfico 2- Resultados do Experimento 2



Fonte: autoria própria

Através dos resultados, é possível perceber uma queda no desempenho do sistema quando se aumenta o número de coeficientes utilizados, para o conjunto composto por todas as palavras. Quando se analisa apenas o conjunto contendo pronúncias dos dígitos, os resultados mantêm-se praticamente invariáveis.

A redução da taxa de acerto com o aumento do número de coeficientes pode estar relacionada com o aumento do número de entradas na rede neural e o consequente aumento na dificuldade da rede neural em extrair o conhecimento sobre o processo.

Além disso, a utilização de um número menor contribui para o desempenho computacional do sistema, uma vez que a dimensão do vetor de entrada da RNA é menor, necessitando menos processamento, se considerada a mesma topologia.

7.3 EXPERIMENTO 3

No terceiro experimento foi analisada a influência da topologia de RNA utilizada nos resultados do sistema de reconhecimento de fala proposto. Foram avaliadas as taxas de acerto para redes neurais com diferentes quantidades de neurônios na camada intermediária.

Os resultados são mostrados na Tabela 4.

Através dos resultados obtidos é possível verificar a dependência do desempenho do sistema com relação à topologia da RNA empregada, quando operando com o conjunto completo de dados. Embora os resultados para o conjunto de dados contendo apenas os dígitos apresentem algumas diferenças, estes valores diferem muito sutilmente, o que não permite garantir que os resultados dependem diretamente da topologia.

Tabela 4 - Resultado: Experimento 3

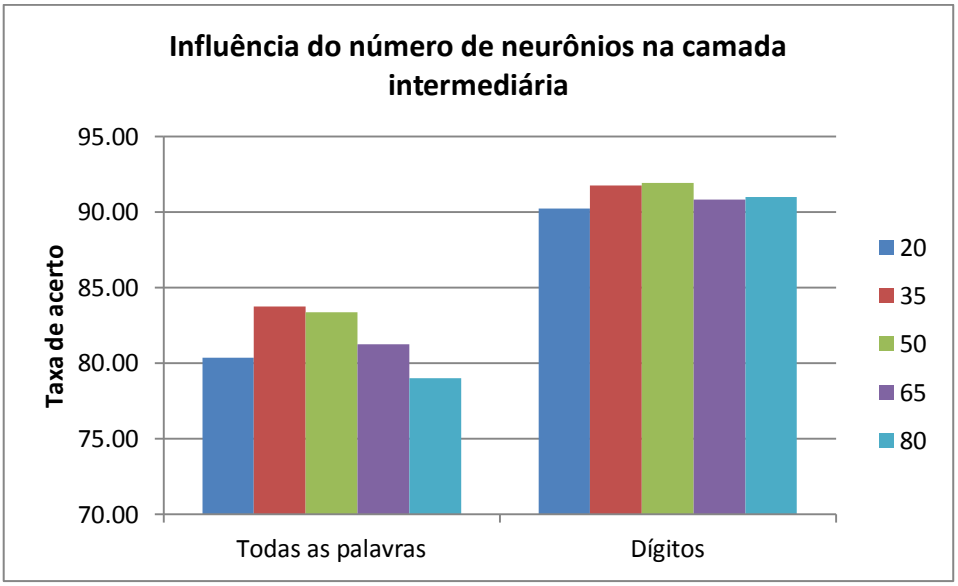
Número de neurônios	Taxa de acerto			
	Todas as palavras		Dígitos	
	Média	Desvio Padrão	Média	Desvio Padrão
20	80.38%	1.92%	90.27%	1.55%
35	83.75%	1.69%	91.78%	1.15%
50	83.39%	3.22%	91.92%	2.64%
65	81.24%	4.58%	90.83%	3.95%
80	79.00%	4.69%	91.00%	4.06%

Fonte: autoria própria.

Os resultados são apresentados também no Gráfico 3.

De qualquer forma, em ambos os conjuntos de dados, as redes neurais com 35 e 50 neurônios na camada intermediária apresentaram os melhores resultados.

Gráfico 3- Resultados do Experimento 3



Fonte: autoria própria.

A redução da taxa de acerto está ligada a dificuldade da rede em extrair conhecimento sobre o processo. No experimento onde foram utilizados apenas 20 neurônios, a reduzida taxa de acerto leva a crer que a RNA foi incapaz de armazenar todo o conhecimento em poucas conexões sinápticas.

Vale notar também o aumento do desvio padrão para taxa de acerto com o aumento do número de neurônios. O que significa uma redução da precisão da rede neural entre os treinamentos. Isso pode estar relacionado, além da maior dificuldade na extração de conhecimento, com a maior dependência dos valores iniciais de pesos sinápticos atribuídos a RNA.

8 CONCLUSÕES

Neste trabalho foi apresentado um sistema de reconhecimento de fala para palavras isoladas, com vocabulário restrito e pequeno, independente de locutor e baseado em redes neurais artificiais.

O sistema descrito foi capaz de classificar comandos de fala incluindo as pronúncias de dígitos (zero, um,... , nove), comandos de direção (esquerda, direita, frente e atrás), comandos de ação (sobe, desce, liga, desliga, acende, apaga, abre e fecha) e ainda, as palavras casa, carro, luz e porta.

No decorrer do trabalho foi apresentada uma abordagem completa do problema de reconhecimento de fala, tratando da produção fisiológica da fala e seu modelo digital, das características dos sons e dos sinais de fala, e descrevendo detalhadamente os processos envolvidos no reconhecimento, como a aquisição do sinal, o pré-processamento, a detecção de palavras, a extração de características e a classificação dos padrões.

A principal contribuição deste trabalho foi a proposta de modificação ao algoritmo de detecção de palavras já proposto na literatura. A modificação incorporou ao algoritmo maior robustez a efeitos indesejáveis gerados no processo de produção da fala, como os sons gerados na boca ao bater ou lamber os lábios e respirar e também robustez a alguns ruídos do ambiente, como por exemplo, o som produzido ao bater a porta da sala, etc.

Entre as técnicas de extração de características, foram apresentadas as principais técnicas utilizadas para representação paramétrica do sinal de fala, enfatizando a técnica de extração dos coeficientes mel-cepstrais, utilizada no desenvolvimento do trabalho.

A técnica de extração dos coeficientes mel-cepstrais foi escolhida por apresentar, de acordo com a literatura, melhores resultados quando comparada as outras técnicas, para a tarefa de classificação de comando de fala.

Para a tarefa de reconhecimento de padrões, optou-se por utilizar as Redes Neurais Artificias, pois se trata de uma técnica recente – quando comparada às cadeias ocultas de Markov, por exemplo – em aplicações de reconhecimento de fala. Além disso, as RNA têm mostrado grande potencial para esta aplicação, conforme vêm mostrando os trabalhos mais recentes.

Para avaliar o desempenho do sistema proposto, foram realizados três experimentos. No primeiro, foi comparado o desempenho da classificação de padrões frente aos algoritmos de detecção de palavras proposto neste trabalho e sua versão original. Através dos experimentos, verificou-se a superioridade da técnica aqui apresentada.

Nos outros experimentos foram avaliadas a influência do número de coeficientes MFCC usados na representação do sinal e de diferentes topologias de redes neurais.

Em todos os casos, foram avaliados os resultados utilizando-se o conjunto completo de dados e um subconjunto, contendo apenas as pronúncias dos dígitos.

De maneira geral, observa-se que os resultados para os testes com o subconjunto dos dígitos apresenta sempre melhores resultados, isso porque, trata de um caso particular do conjunto de dados, com um conjunto menor de classes.

Os melhores resultados apresentaram taxa de acerto de 83.75% de acerto para o conjunto completo de dados, utilizando-se 12 coeficientes MFCC para cada segmento do sinal de fala e uma rede neural com 35 neurônios na camada escondida.

Para os testes com o subconjunto dos dígitos, o melhor resultado obtido foi de 92,09% de acerto, no experimento onde foram utilizados 18 coeficientes mel-cepstrais e uma rede neural com 50 neurônios na camada intermediária.

As taxas médias de acerto entre todos os experimentos realizados ficaram em torno de 81,8% e 91,4% para o conjunto completo de palavras e para o subconjunto dos dígitos, respectivamente.

Além disso, um dos principais resultados deste trabalho foi a geração do banco de dados. O banco de dados foi gravado com 30 estudantes da Universidade do Estado de Santa Catarina, distribuídos em 66.7% do sexo masculino e 33.3% do sexo feminino, e com naturalidade de diversas cidades da região sul do país, de acordo com a metodologia descrita no trabalho.

A gravação deste banco de dados permitirá a continuidade dos trabalhos relacionados ao reconhecimento de fala. Pretende-se ainda, disponibilizar este banco de dados, através da internet, à outros pesquisadores do tema.

Como sugestão de trabalhos futuros, pode-se considerar a avaliação do desempenho do algoritmo de detecção de palavras proposto neste trabalho com as outras técnicas, brevemente discutidas, baseadas em energia e taxa de cruzamento

por zero. É possível ainda, considerar a inclusão da ZCC no próprio algoritmo de detecção de palavras ou a inclusão de parâmetros como energia e a ZCC na descrição paramétrica do sinal de fala utilizada como entrada pela rede neural.

Por fim, sugere-se a implementação dos algoritmos aqui utilizados em dispositivos portáteis e embarcados. O conjunto de palavras escolhidos durante o trabalho dão margem para aplicações em automação residencial, tecnologias assistivas aplicadas à eletrodomésticos ou sistema de telefonia.

9 REFERÊNCIAS

AHMED, N.; NATARAJAN, T.; RAO, K. R. Discrete Cosine Transform. **IEEE Transactions on Computers**, v. C-23, n. 1, p. 90-93, January 1974.

AIDA-ZADE, K. R.; ARDIL, C.; RUSTAMOV, S. S. Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems. **International Journal of Signal Processing**, v. 3, p. 105–111, 2006.

BLINN, J. F. What's that deal with the DCT? **IEEE Computer Graphics and Applications**, v. 13, n. 4, p. 78-83, July 1993.

BOURLARD, H.; MORGAN, N. Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions. In: GILES, C. L.; GORI, M. **Adaptive Processing of Sequences and Data Structures**. Berlin: Springer-Verlag, v. 1387, 1998. p. 389-417.

BRESOLIN, A. D. A. **Reconhecimento de voz através de unidades menores do que a palavra, utilizando Wavelet Packet e SVM, em uma nova estrutura hierárquica de decisão**. Universidade Federal do Rio Grande do Norte. Rio Grande do Norte. 2008.

BROOKES, M. **VOICEBOX:** Speech Processing Toolbox for MATLAB, London, 2005. Disponivel em: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>. Acesso em: 9 Jun. 2012.

COMBRINCK, H. P.; BOTHA, E. C. **On The Mel-scaled Cepstrum**. Proceedings of the Seventh Annual South African Workshop on Pattern Recognition. University of Cape Town: Cape Town. 1996.

COOLEY, J. W.; TUKEY, J. W. An Algorithm for the Machine Calculation of Complex Fourier Series. **Mathematics of Computation**, 19, n. 90, 1965. 297.

DAVIS, K.; BIDDULPH, R.; BALASHEK, S. Automatic Recognition of Spoken Digits. **Journal of the Acoustical Society of America**, 24, n. 6, 1952. 637-642.

DAVIS, S. B.; MERMELSTEIN, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. **IEEE Transactions on Acoustic, Speech, and Signal Processing**, v. 28, n. 4, p. 357-366, August 1980.

DELLER, J. R.; HANSEN, J. H. L.; PROAKIS, J. G. **Discrete-Time Processing of Speech Signals**. New York: Macmillan Pub. Co., 1993.

DEMUTH, H.; BEALE, M.; HAGAN, M. **Neural Network Toolbox 6:** User's Guide. Natick: The MathWorks, Inc., 2010.

FURUI, S. 50 years of progress in speech and speaker recognition. **Proceedings of the 10th International Conference on Speech and Computer**, Patras, 2005. 1-9.

GOLD, B.; MORGAN, N. **Speech and Audio Signal Processing:** Processing and Perception of Speech and Music. New York: John Wiley & Sons Inc, 1999. ISBN 0-471-35154-7.

HAYKIN, S. **Redes Neurais:** princípios e prática. Tradução de Paulo Martins Engel. 2ª. ed. Porto Alegre: Bookman, 2001.

HESS, W. **Pitch Determination of Speech Signals**. New York: Springer-Verlag, 1983.

HOWARD, I. S. **Speech fundamental period estimation using pattern classification**. University of London. London. 1992.

HUNT, A. What is speech recognition? **comp.speech Frequently Asked Questions WWW**, 05 Setembro 1997. Disponivel em: <http://www.speech.cs.cmu.edu/comp.speech/Section6/Q6.1.html>. Acesso em: 26 Novembro 2011.

JUANG, B. H.; RABINER, L. R. Automatic Speech Recognition – A Brief History of the Technology Development. In: BROWN, K. **Elsevier Encyclopedia of Language and Linguistics**. Second Edition. ed. [S.l.]: Elsevier, 2005. p. 806-819. ISBN 978-0-08-044854-1.

KEERIO, A. et al. On Preprocessing of Speech Signals. **International Journal of Signal Processing**, v. 5, n. 3, p. 216-222, February 2009.

KESARKAR, M. P. **Feature Extraction for Speech Recognition**. Indian Institute of Technology. Bombay. 2003.

LAMEL, L. F. et al. An Improved Endpoint Detector for Isolated Word Recognition. **IEEE Transactions on Acoustics, and Signal Processing**, v. ASSP-29, n. 4, p. 777-785, August 1981.

LATHI, B. P. **Linear Systems and Signals**. 2ª. ed. New York: Oxford University Press, 2004.

LIU, J.; SUN, J.; WANG, S. Pattern Recognition: An overview. **International Journal of Computer Science and Network Security**, v. 6, n. 6, p. 57-61, June 2006.

MAGALHÃES, H. A. Pré-ênfase do Sinal de Voz para Extração de Formantes. **CEFALA - Centro de Estudos da Fala, Acústica, Linguagem e Música**, 2002. Disponivel em: <http://www.cefala.org/~hermes/pubs/music/prnfase.htm>. Acesso em: 9 Dezembro 2011.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, 5, 1943. 115-133.

NETO, J. A. O.; CASTRO, M. A. A.; FELIX, L. B. Reconhecimento de Comandos de Voz para o Acionamento de Cadeira de Rodas. **Anais do XVIII Congresso Brasileiro de Automática**, Bonito, MS, 2010. 3819-3824.

NUNES, H. F. **Reconhecimento de Fala baseado em HMM**. Universidade Estadual de Campinas. Campinas. 1996.

OLIVEIRA, D. D. H. **Fonética e Fonologia**. Universidade Federal da Paraíba. Paraíba.

PERETTA, I. S. et al. Reconhecimento de Comandos de Voz baseados em Filtro Wavelet utilizando Redes Neurais Artificiais. **Anais do IX Congresso Brasileiro de Redes Neurais / Inteligência Computacional**, Ouro Preto, 2009.

PICONE, J. W. Signal modeling techniques in speech recognition. **Proceedings of the IEEE**, v. 81, n. 9, p. 1215-1247, 1993.

PROAKIS, J. G.; MANOLAKIS, D. G. **Digital Signal Processing:** Principles Algorithms, and Applications. 3ª. ed. Upper Saddle River, New Jersey: Prentice Hall, 1996.

RABINER, L. et al. Speaker-independent recognition of isolated words using clustering techniques. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, 27, n. 4, Agosto 1979. 336-349.

RABINER, L. R.; SAMBUR, M. R. An Algorithm for Determining the Endpoints of Isolated Utterances. **The Bell System Technical Journal**, v. 54, n. 2, p. 297-315, February 1975.

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

RABINER, L.; JUANG, B.-H. **Fundamentals of Speech Recognition**. Englewood Cliffs: Prentice-Hall, 1993.

RIEDMILLER, M.; BRAUN, H. **A Direct Adaptive Method for Faster Backpropagation Learning:** The RPROP Algorithm. IEEE International Conference on Neural Networks. San Francisco: [s.n.]. 1993. p. 586-591.

RUDNICKY, A. I.; HAUPTMANN, A. G.; LEE, K.-F. Survey of current speech technology, New York, 37, n. 3, Março 1994.

RUMELHART , D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Internal Representations by Error Propagation. In: RUMELHART, D. E.; MCCLELLAND, J. L. **Parallel Distributed Processing:** Explorations in the Microstructure of Cognition. Cambridge: Bradford Books/MIT Press, 1986.

RUNSTEIN, F. O. **Sistema de Reconhecimento de Fala Baseado em Redes Neurais Artificiais**. UNICAMP. Campinas. 1998.

RUNSTEIN, F.; VIOLARO, F. **An isolated-word speech recognition system using neural networks**. 38th Midwest Symposium on Circuits and Systems. Proceedings. [S.l.]: [s.n.]. 1996. p. 550-553.

SÁ, J. P. M. D. **Pattern Recognition - Concepts Methods and Applications**. Berlin: Springer, 2001.

SABAH, R.; AINON, R. N. **Isolated Digit Speech Recognition in Malay Language Using Neuro-Fuzzy Approach**. Third Asia International Conference on Modelling & Simulation. [S.l.]: [s.n.]. 2009. p. 336-340.

SAHA, G.; CHAKRABORTY, S.; SENAPATI, S. A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications. **Proceedings of the National Conference on Communications, NCC-2005**, Kharagpur, Índia, p. 291-295, January 2005.

SAMOUELIAN, A. **Isolated voiced digit recognition using inductive inference**. Proceedings of Digital Processing Applications (TENCON '96). [S.l.]: [s.n.]. 1996. p. 119-124.

SILVA, A. H. P. **Língua Portuguesa I:** fonética e fonologia. Curitiba: IESDE Brasil S.A, 2007.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes Neurais Artificiais para engenharia e ciências aplicadas**. 1. ed. São Paulo: Artliber, 2010.

STEVENS, S. S.; VOLKMAN, J.; NEWMAN, E. B. A Scale for Measurement of the Psychological Magnitude Pitch. **Journal of the Acoustical Society of America**, v. 8, p. 185-190, January 1937.

SUK, S.-Y.; KOJIM, H. Voice Activated Appliances for Severely Disabled Persons. In: MIHELIC, F.; ZIBERT, J. **Speech Recognition:** Technologies and Applications. [S.l.]: In-Tech Education and Publishing, 2008. ISBN 978-953-7619-29-9. Available from: http://www.intechopen.com/articles/show/title/.

SUN, X. **A pitch determination algorithm based on subharmonic-to-harmonic ratio**. Proceedings of International Conference on Spoken Language Processing. Beijing: [s.n.]. 2000. p. 676-679.

TEBELSKIS, J. **Speech recognition using neural networks**. Carnegie Mellon University. Pittsburg, Pensilvania. 1995.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. 4ª. ed. [S.l.]: Elsevier/Academic Press, 2008.

TRENTIN, E.; GORI, M. A survey of hybrid ANN/HMM models for automatic speech recognition. **Neurocomputing**, 37, n. 1-4, Abril 2001. 91-126.

YOMA, N. B. **Reconhecimento Automático de Palavras Isoladas: Estudo e Aplicação dos Métodos Determinísticos e Estocásticos**. Universidade Estadual de Campinas. Campinas. 1993.

ZHENG, N. **Feature Extraction in Automatic Speech Recognition**. The Chinese University of Hong Kong. Hong Kong, p. 30. 2007.

ZWICKER, E.; TERHARDT, E. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. **Journal of the Acoustical Society of America**, v. 68, n. 5, p. 1523-1525, December 1980.

10 APÊNDICE A – ARTIGO ACEITO PARA PUBLICAÇÃO NO XIX CONGRESSO BRASILEIRO DE AUTOMÁTICA – CBA 2012

RECONHECIMENTO DE FALA POR MEIO DE UM NOVO ALGORITMO DE DETECÇÃO DE PALAVRA E DE REDES NEURAIS ARTIFICIAIS

Guilherme M. Zilli*, Lucas H. Negri*, Aleksander S. Paterno*

*Departamento de Engenharia Elétrica, Universidade do Estado de Santa Catarina
Campus Universitário Prof. Avelino Marcante s/n, Bom Retiro, 89223-100, Joinville, SC, Brasil

Emails: guilherme.m.zilli, lucashnegri@gmail.com, dee2asp@joinville.udesc.br

Abstract— This paper describes an speaker-independent automatic speech recognition system for isolated words and a small vocabulary. This work introduces a modification to an algorithm already found in literature for word detection and describes the techniques used for the signal pre-processing and features extraction. A multilayer perceptron artificial neural network was employed for the final classification task, trained with the resilient backpropagation algorithm. In the performed experiments the modified algorithm, proposed in this work, shown better classification rates than the original algorithm.

Keywords— Speech Recognition, Mel Frequency Cepstral Coefficient, Artificial Neural networks, iRPROP.

Resumo— Este artigo descreve um sistema de reconhecimento automático de fala para palavras isoladas, com vocabulário restrito e independente de locutor. Neste trabalho apresenta-se uma modificação para um algoritmo já encontrado na literatura para a detecção de palavras e descreve as técnicas utilizadas para pré-processamento e extração de características dos sinais. Uma rede neural artificial do tipo perceptron de múltiplas camadas foi utilizada na tarefa final de classificação, treinada pelo algoritmo resilient backpropagation. Nos experimentos realizados o algoritmo modificado, proposto neste trabalho, resultou em taxas de classificação melhores do que o algoritmo original.

Palavras-chave— Reconhecimento de Fala, Coeficiente Mel-Cepstral, Redes Neurais Artificiais, iRPROP.

1 Introdução

Os sistemas de reconhecimento automático de fala têm por objetivo extrair informações de um sinal de fala, processá-lo e classificá-lo de acordo com seu conteúdo e contexto. Estes sistemas constituem-se basicamente de três subsistemas: de aquisição e pré-processamento do sinal de fala, de extração de características e de classificação/reconhecimento de padrões.

Na etapa de pré-processamento é feita a conformação espectral do sinal de fala e a detecção dos limites do início e final da palavra. Na extração de características, o sinal de fala é transformado em um conjunto de parâmetros, que representam as características mais relevantes do sinal. Por último, a etapa de reconhecimento de padrões classifica estes conjuntos de características, determinando qual palavra foi pronunciada.

O desenvolvimento e o aprimoramento dos algoritmos de extração de características e dos algoritmos de reconhecimento de padrões, bem como, a utilização de inteligência computacional, têm aproximado cada vez mais o desempenho dos sistemas de reconhecimento de voz digitais do sistema humano. A evolução dos sistemas de reconhecimento automático de fala, bem como os principais projetos e instituições no qual tais projetos foram desenvolvidos são descritos em (Juang and Rabiner, 2005; Gold and Morgan, 2000; Furui, 2005).

O reconhecimento automático de fala pode ser classificados de acordo com diferentes aspectos, sendo os mais comuns a dependência de locutor, o tamanho do vocabulário tratado e o estilo de fala (Nunes, 1996; Tebelskis, 1995).

Neste trabalho é apresentado um sistema de reconhecimento automático de palavras isoladas, independente de usuário e com vocabulário restrito utilizando redes neurais artificiais (RNAs) e coeficientes mel-cepstrais. O artigo propõe também uma modificação no algoritmo de detecção de palavras apresentado em (Bresolin, 2008) para o tratamento de problemas apontados na literatura (Lamel et al., 1981).

O algoritmo de detecção de palavra, também chamado de (*end-point detection*, tem por objetivo isolar a pronúncia da palavra do silêncio ou ruído de fundo que precedem e sucedem a pronúncia durante a gravação. A modificação proposta neste trabalho visa tornar a detecção de palavras mais robusta às interferências geradas, principalmente, durante a produção da fala.

O trabalho está dividido da seguinte forma: na Seção 2 são apresentadas as características do banco de dados criado para o desenvolvimento do trabalho, os algoritmos utilizados para pré-processamento e extração de características e aspectos gerais das RNAs utilizadas; os resultados dos experimentos são expostos na Seção 3 e discutidos na Seção 4; as conclusões e sugestões para trabalhos futuros são apresentadas na Seção 5.

2 Metodologia

Como descrito anteriormente, os sistemas de reconhecimento de fala são formados pelos subsitemas: de aquisição e pré-processamento do sinal de fala, de extração de características e de classificação de

padrões.

A etapa de aquisição e pré-processamento é a parte inicial do processo, onde são feitas a aquisição e digitalização do sinal de áudio e sua posterior filtragem (para eliminar interferências provenientes da rede elétrica ou limitar a banda espectral, por exemplo), além da etapa de *end-point detection* que consiste na determinação do início e final da palavra pra que possa ser extraída do sinal gravado.

Na sequência, a etapa de extração de características, ou modelagem do sinal, permite reduzir a dimensão do conjunto de dados dos sinais de áudio, resumindo-o e excluindo informações desnecessárias e/ou redundantes. Existe uma grande quantidade de representações paramétrica do sinal de fala, incluindo energia de um curto intervalo de tempo, taxa de cruzamento por zero, entre outros (Rabiner and Juang, 1993). No entanto, os métodos de análise espectral são os mais importantes. Entre os métodos de análise espectral encontram-se diversas técnicas, como, banco de filtros, banco de filtros digitais, coeficientes melcepstrais e coeficientes de predição linear, sendo que estes dois últimos destacam-se em aplicações de reconhecimento de fala (Picone, 1993).

Por último, a etapa de reconhecimento/classificação de padrão consiste basicamente na comparação dos dados, resultantes da etapa anterior, com informações previamente conhecidas, relacionadas às palavras que o sistema é capaz de identificar. Até meados da década de 80 as principais técnicas de reconhecimento de padrão eram baseadas em métodos para como a *Dynamic Time Warp* e modelos ocultos de Markov, sendo que nenhuma outra grande técnica foi desenvolvida até este período (Gold and Morgan, 2000). A partir do trabalho de (Rumelhart et al., 1985), muitas aplicação de redes neurais artificiais em reconhecimento de fala começaram a surgir. Segundo (Tebelskis, 1995), as primeiras aplicações consistiam em tarefas simples, como classificação dos sons em vocálicos ou não vocálicos ou classificação entre sons fricativos, nasais ou plosivos. O sucesso nestas aplicações motivou os pesquisadores a utilizar redes neurais artificiais em reconhecimento de fonemas e, em seguida, de palavras.

O sistema proposto deve reconhecer comandos de fala como dígitos (*zero*, *um*, ... , *nove*), comandos de direção (*esquerda*, *direita*, *frente* e *atrás*), comandos de ação (*sobe*, *desce*, *liga*, *desliga*, *acende*, *apaga*, *abre* e *fecha*) e as palavras *casa*, *carro*, *luz* e *porta*.

2.1 Gravação do banco de dados

A base de dados para treinamento, teste e validação do sistema foi gravada a parir de um computador pessoal, usando uma placa de áudio IDT High Definition Audio e um microfone Microsoft® LifeChat™ LX-2000. Durante as gravações o microfone foi mantido estático, a uma distância entre 15 e 20 cm do usuário. As gravações foram feitas por meio do software Audaticy 1.3, a uma frequência de amostragem de 11025 Hz, resolução de 16 bits, duração de um segundo, e codificadas em PCM - *Pulse Code Modulation* -, sem comprensão. Nenhum tratamento do sinal (e. g., filtragem, remoção do ruído) foi feita na etapa de gravação da base de dados.

O ambiente no qual foram realizadas as gravações foi uma sala com ruído ambiente, gerado principalmente por aparelho de ar-condicionado e computadores.

Utilizou-se um total de 30 usuários, sendo estes divididos em 20 usuários do sexo masculino e 10, do sexo feminino. Os usuários possuíam naturalidade de diversas cidades da região sul do Brasil, apresentando, assim, diferentes sotaques.

Foram gravadas as elocuções das 26 palavras, pronunciadas 5 vezes cada uma, para cada um dos usuários, totalizando 3900 arquivos de áudio.

2.2 Pré-processamento

Após a gravação da base de dados, deu-se início aos algoritmos que constituem o sistema de reconhecimento automático de fala.

Na primeira etapa, que consiste no pré-processamento, foi feita a filtragem e pré-ênfase dos sinais. Essa etapa é também chamada de conformação espectral (Picone, 1993).

A fim de reduzir efeitos das interferências geradas, principalmente, pela rede de energia elétrica (interferência audível de 60 Hz), os sinais gravados foram filtrados por um filtro FIR digital passa-banda, com banda de passagem entre 100 Hz e 4000 Hz. Na sequência, o sinal foi filtrado por um filtro de pré-ênfase. O filtro de pré-ênfase é um filtro FIR de ordem unitária, como pode ser visto na Equação 1:

$$H_{pre}(z) = 1 + \alpha_{pre} z^{-1} \tag{1}$$

onde α_{pre} apresenta-se tipicamente entre $-0,4$ e 1. Neste trabalho, utilizou-se $\alpha_{pre} = 0,95$.

Segundo (Picone, 1993) as duas principais vantagens da pré-ênfase são: compensar a atenuação natural, de aproximadamente 20 dB por década, devido as características fisiológicas do sistema de produção da fala e evidenciar componentes de frequência a partir de 1 kHz, faixa em que o sistema auditivo é mais sensível.

Após a conformação espectral foi feito o *endpoint detection*. O *endpoint detection* (Lamel et al., 1981) consiste na detecção dos pontos de início e final da pronúncia da palavra presente em uma gravação. Esta detecção permite extrair apenas o intervalo onde há sinal de fala, excluindo os intervalos em que há apenas ruído de fundo.

Ainda segundo o autor, a extração da palavra é um processo muito importante, pois reduz o processamento do sinal e está diretamente relacionado à precisão do sistema de reconhecimento de fala.

Grande parte dos sistemas de reconhecimento de fala utilizam características como energia e taxa de cruzamento por zero para determinação dos limiares das palavras. Um algoritmo baseado nestas características é apresentado em (Rabiner and Schafer, 1978).

Neste trabalho, foi utilizada uma adaptação do algoritmo proposto por (Bresolin, 2008). Na proposta original do autor, o algoritmo calcula a mediana da energia do sinal, que representa aproximadamente a energia do sinal de fundo, e considera como limiar de detecção o valor de três vezes esta mediana. O sinal é segmentado em janelas de 5 ms e a energia de cada janela é medida. Considera-se como início da palavra o instante em que a energia do segmento é superior ao limiar pela primeira vez, e o final, quando a energia do segmento for menor que o limiar pela última vez.

A principal modificação proposta neste trabalho foi a consideração de um limiar temporal, ou seja, é necessário que a energia do sinal seja maior que o limiar de detecção durante um período equivalente à 10 janelas (aproximadamente 36 ms), para que seja considerado o início da palavra. De maneira análoga ocorre o processo para determinação do final da palavra. Além disso, neste trabalho foram utilizados janelas de tamanho 3, 6 ms e quatro vezes o valor da mediana.

A consideração desse limiar temporal permite reduzir problemas apontados por (Lamel et al., 1981), como o ruído gerado pela respiração, pela batida e estalos gerados pela boca e pela abertura e fechamento lábios.

2.3 Extração de características

A função da extração de caraterística é representar os eventos acústicos relevantes no sinal de fala em termos de um conjunto eficiente e compacto de parâmetros (Rabiner and Schafer, 1978; Juang and Rabiner, 2005).

A extração de coeficientes de frequência mel-cepstrais (*Mel Frequency Cepstral Coefficient*, MFCC) e de coeficientes de predição linear é empregada na maioria dos sistemas de reconhecimento de fala (Gold and Morgan, 2000). No entanto, alguns trabalhos mostram maior eficiência em sistemas que empregam características mel-cepstrais (Zade et al., 2006; Lévy et al., 2003).

A essência dos coeficientes mel-cepstrais é derivada da modelagem da geração da fala. A modelagem completa da produção de fala pode ser encontrada em (Rabiner and Schafer, 1978). Um modelo simplificado pode ser encontrado em (Zade et al., 2006). No modelo simplificado, que pode ser visualizado na Figura 1, o autor sugere que

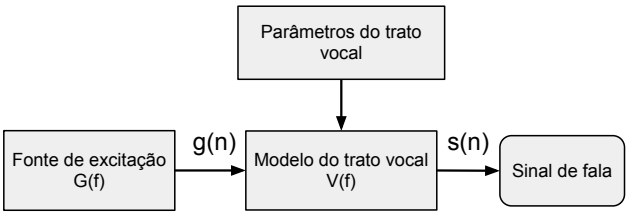


Figura 1: Modelo da produção da fala.

a geração da fala consiste na geração do sinal de excitação e na filtragem pelo trato vocal.

A separação destas duas componentes (sinal de excitação e filtro) permite obter um modelo do trato vocal. Sabendo que a filtragem consiste na operação de convolução (Equação 2), tem-se:

$$s(n) = g(n) * v(n) \tag{2}$$

onde $s(n)$ representa o sinal de fala, $g(n)$ denota a fonte de excitação e $v(n)$ a reposta ao impulso do trato vocal. Conhecendo as propriedades da convolução, podemos representar a Equação 2 no domínio da frequência, como visto na Equação 3.

$$S(f) = G(f)V(f) \tag{3}$$

Aplicando o logaritmo em ambos os lados da Equação 3 obtemos a Equação 4:

$$\begin{aligned} \log(S(f)) &= \log(G(f)V(f)) \\ &= \log(G(f)) + \log(V(f)) \end{aligned} \tag{4}$$

sendo que desta forma é possível separar o sinal de excitação e o modelo do trato vocal. Os coeficientes cepstrais são calculados por meio da transformada inversa de Fourier do logaritmo do espectro, como visto na Equação 5.

$$c(n) = \frac{1}{N_s} \sum_{k=0}^{N_s-1} \log(|S(k)|) e^{2\pi N_s k n} \tag{5}$$
$$0 \leq n \leq N_s - 1$$

Os termos de baixa ordem dos coeficientes cepstrais estão relacionados com a forma do trato vocal, enquanto os temos de ordens mais elevadas, com o sinal de excitação. Na prática, para reconhecimento de fala, apenas os termos de baixa ordem ($n < 20$) são utilizado (Picone, 1993).

O sinal de voz não é estacionário no tempo. Para que se possa analisar o sinal de voz pode-se, então, recorrer à segmentação do sinal e, nesse caso, considerá-lo quasi-estacionário por partes.

A duração dos segmentos está diretamente relacionada com a dinâmica do trato vocal. Para sinais de fala são usadas janelas de duração entre 10 ms e 40 ms (Picone, 1993). Aplica-se a estes segmentos uma janela de ponderação, como as janelas de Hanning ou Hamming, com o objetivo

de suavizar as variações espectrais do sinal. Também por esta razão, é comum sobrepor parte da janela anterior à janela na qual são extraídas as características. Assim, cada janela é formada por parte dos dados da janela anterior e novos dados do sinal de fala. A relação entre a quantidades de dados antigos e o tamanho da janela define a razão de sobreposição, que em geral, está entre 30 % e 60 %.

Desta forma, a extração dos coeficientes MFCC deve ser feita quadro por quadro. Assim, em cada segmento extrai-se um conjunto de coeficientes MFCC ($c_i(n) = c(0), c(1), ..., c(p)$), onde $c(n)$ são os coeficientes MFCC, p é a ordem do extrator ou o número de coeficientes extraídos, e i é o índice do segmento do qual os coeficientes foram extraídos.

Ao final do processo de extração de características têm-se, então, um vetor de características, formados pelos coeficientes extraídos de cada segmento ($v(k) = c_1(n), c_2(n), ..., c_k(n)$). Este vetor é usado, posteriormente, na etapa de reconhecimento de padrões.

Como o sinal de voz possui diferentes durações, é natural que a dimensão do vetor de características esteja associada à duração do áudio. No entanto, quando se utiliza um classificador de padrões com um número fixo de entradas (como é o caso das redes neurais artificiais) é necessário que a dimensão do vetor de características seja compatível com a dimensão da entrada do classificador.

Para solucionar este problema, definiu-se um número fixo de segmentos (N_{seg}) para extração de característica e, então, o tamanho das janelas foi ajustado dinamicamente, de modo que, para cada palavra, o tamanho das janelas era suficiente para que a pronúncia fosse completamente descrita em N_{seg} segmentos. A escolha de N_{seg} está condicionada ao fato de que o tamanho da janela garanta a quase-estacionaridade do segmento, ou seja, tenha duração inferior a 40 ms.

Neste trabalho utilizaram-se os algoritmos presentes no toolbox *Voicebox: Speech Processing Toolbox for MATLAB* para extração dos coeficientes MFCC. Foram extraídos 12 coeficientes para cada segmento. Utilizou-se tamanho dos segmentos variável, de modo que, todo o sinal correspondente à pronúncia da palavra fosse dividido em 20 quadros. Além disso, foram utilizadas janelas de Hamming, com sobreposição de $33,3$ % entre janelas consecutivas.

2.4 Classificação por meio de redes neurais artificiais

A etapa final do sistema de reconhecimento de fala consiste no reconhecimento/classificação dos padrões, a partir da apresentação do vetor de características. Neste trabalho foram utilizadas redes neurais artificiais para a classificação dos comandos de fala.

Uma RNA é a implementação de um modelo computacional formado por um conjunto de processadores simples, os neurônios, sendo que seu poder computacional emerge da estrutura formada pela interconexão destes neurônios. Este modelo computacional conexionista é comumente utilizado em problemas que envolvem a busca por padrões em conjuntos de dados, como é o caso dos problemas de classificação (Haykin, 2001).

Dentre os modelos de RNAs utilizados para tarefas de classificação, encontram-se as redes perceptron de múltiplas camadas (MLP), que são compostas por uma camada de neurônios para entrada de dados, uma ou mais camadas internas ou escondidas e uma camada de saída (Silva et al., 2010). O treinamento de redes do tipo MLP é comumente realizado por algoritmos de treinamento supervisionados de descida de gradiente, que calculam o gradiente de erro por meio da retropropagação do erro na camada de saída para as camadas anteriores. Entre tais algoritmos pode-se citar o algoritmo que leva o nome de retropropagação de erro (*backpropagation*) e suas modificações como o *resilient backpropagation*(RPROP) (Riedmiller and Braun, 1993) e sua melhoria (iRPROP) (Igel and Hüsken, 2000).

Neste trabalho realiza-se o reconhecimento de palavras (para um vocabulário restrito) por meio de uma rede MLP desenvolvida para classificar as elocuções quanto à palavra pronunciada, sendo que a rede MLP é treinada pelo algoritmo iRPROP. No algoritmo iRPROP os pesos sinápticos são atualizados com passos individuais, ao invés de um passo global, sendo que estes passos são controlados pelo sinal das derivadas parciais do erro em relação ao respectivo peso sináptico, independentemente do valor absoluto das derivadas parciais. Este processo é realizado com o objetivo de acelerar o treinamento da RNA em relação à retropropagação de erro original, e, por realizar o ajuste automático dos passos de aprendizagem, exige menos ajustes manuais de parâmetros (Riedmiller and Braun, 1993; Igel and Hüsken, 2000).

2.5 Experimento de classificação de palavras

Utilizou-se as técnicas previamente descritas nesta Seção em conjunto para classificar elocuções de acordo com a palavra pronunciada. As elocuções foram gravadas, pré-processadas e passaram por um método de extração de características que resultou em 240 coeficientes reais para cada elocução.

Uma RNA do tipo MLP foi desenvolvida para esta tarefa de classificação, tendo como entrada os 240 coeficientes extraídos das elocuções e tendo como saída 26 neurônios, além de uma camada escondida com 240 neurônios. Utilizou-se uma co-

dificação de saída do tipo *one of C-classes* (Silva et al., 2010), onde cada neurônio está relacionado à uma classe. Assim, a classe (palavra) relacionada ao neurônio na última camada com o maior valor de saída é tida como a resposta da rede para uma classificação. A RNA foi treinada pelo algoritmo iRPROP, utilizando 3/5 do conjunto de dados. O restante dos dados foi utilizado unicamente para validação da RNA. Como cada pessoa repetiu a mesma palavra 5 vezes, sendo que 3 repetições foram utilizadas no treinamento e 2 na validação. Utilizou-se um número fixo de 200 épocas para o treinamento.

A RNA utilizou funções de ativação sigmoidais simétricas, sendo que os dados foram propriamente ajustados à faixa de operação das funções de ativação utilizadas.

Outra RNA foi desenvolvida seguindo a mesma configuração da primeira, porém utilizando somente 200 neurônios na camada intermediária e somente 10 neurônios de saída. Esta RNA foi treinada e validada somente para os dados referentes as elocuções de dígitos (zero até nove). O número de neurônios nas camadas intermediárias das RNAs desenvolvidas foi ajustado empiricamente.

Com o objetivo de avaliar a modificação realizada no algoritmo de detecção de palavras, o experimento proposto foi realizado duas vezes, sendo que na primeira vez utilizou-se o algoritmo original e na segunda vez o algoritmo modificado na preparação das entradas para as RNAs, ou seja, 4 diferentes RNAs foram treinadas no total. A inicialização dos pesos sinápticos das RNAs foi feita de forma pseudo-aleatória, e portanto todos os treinamentos foram realizados 32 vezes para cálculo da média e desvio padrão amostral.

3 Resultados

Na Figura 2 pode-se observar a influência da adaptação proposta na determinação dos limites das palavras. Nesta Figura, mostra-se a forma de onda relativa à uma pronúncia da palavra *frente*, onde a linha pontilhada mostra os limites obtidos pelo algoritmo original e a linha sólida mostra os limites obtidos com o algoritmo modificado.

Na Tabela 3 pode-se visualizar as maiores taxas de acerto (média ± desvio padrão amostral) obtidas no treinamento das RNAs, para o conjunto de dados de treinamento e para o de validação. As RNAs cujos dados foram preparados com o algoritmo de detecção de palavras original são marcadas pela etiqueta *original*, enquanto que as RNAs que utilizaram os dados preparados pelo algoritmo proposto neste trabalho são marcadas pela etiqueta *proposto*. Também evidenciou-se o conjunto de dados como *conjunto completo* ou *somente dígitos*.

Na Figura 3 pode-se visualizar um exemplo do progresso das taxas de classificação para o conjunto de treinamento e de validação durante o treinamento de uma RNA, para o conjunto completo e o algoritmo modificado.

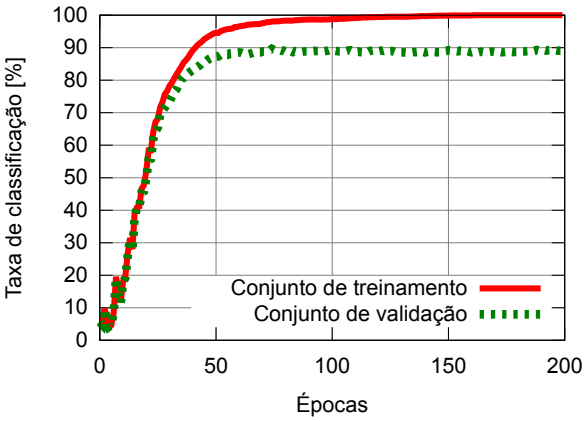


Figura 3: Exemplo do progresso das taxas de classificação para o conjunto de treinamento e de validação durante o treinamento de uma RNA, para o conjunto completo e o algoritmo modificado.

3.1 Discussão

O resultado da computação dos limites de uma pronúncia, visto na Figura 2, mostra que o algoritmo proposto trata dos tipos de problema apontados previamente na literatura (Lamel et al., 1981) ao prover um meio para ignorar a parte inicial do sinal que não contém informação útil para uso posterior.

Como visto na Tabela 3, todas as RNAs conseguiram classificar corretamente todos os dados de treinamento em todos os testes realizados, fato visto pela média de acerto igual a 100 % e desvio padrão igual a 0 %. Contudo, estas RNAs resultaram em erros de classificação na etapa de validação, sendo que as redes treinadas para o número menor de classes (somente dígitos) resultaram em uma maior taxa média de acerto para os testes de validação, como se é esperado devido à esta simplificação do problema.

Na Tabela 3 também torna-se evidente a partir das taxas médias de classificação e dos respectivos desvios que a modificação proposta melhorou as taxas de classificação, tanto para o conjunto de dados completo quanto para o conjunto de dados contendo somente dígitos, sendo que a maior melhoria ocorreu nos experimentos com o conjunto de dados completo. Esta melhoria na taxa de classificação evidencia que o corte da informação inútil na parte inicial do sinal, como visto na Figura 2, que é realizado pela modificação proposta neste trabalho é benéfica.

Algumas aplicações podem utilizar informações de contexto para reduzir as classes possíveis durante o reconhecimento das palavras. Por

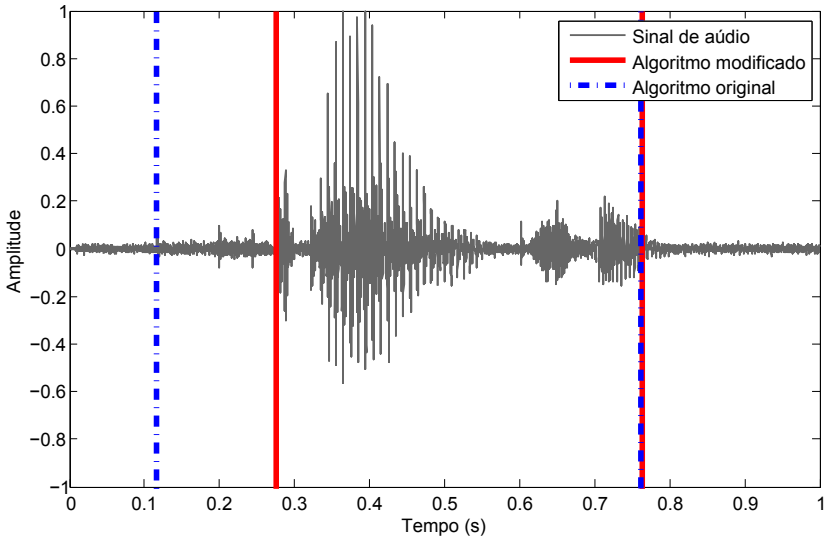


Figura 2: Resultado dos limites computados pelo algoritmo proposto na literatura (Bresolin, 2008) e o algoritmo proposto neste trabalho.

RNA	Treinamento	Validação
Original – Conjunto completo	$100\,\% \pm 0,00\,\%$	$82,7\,\% \pm 0,59\,\%$
Proposto – Conjunto completo	$100\,\% \pm 0,00\,\%$	$89,54\,\% \pm 0,50\,\%$
Original – Somente dígitos	$100\,\% \pm 0,00\,\%$	$91,68\,\% \pm 0,52\,\%$
Proposto – Somente dígitos	$100\,\% \pm 0,00\,\%$	$94,26\,\% \pm 0,53\,\%$

Tabela 1: Estatísticas para o treinamento e validação das RNAs utilizada no experimento para comparação entre o algoritmo de detecção de palavras original e o proposto.

exemplo, em algumas aplicações nenhum comando de voz é iniciado por um dígito, logo seria possível ignorar o resultado dos neurônios referentes as classes não relevantes para o momento. Este uso do contexto é uma técnica relevante que pode ser avaliada futuramente para a melhoria das taxas de classificação.

4 Conclusões

Neste trabalho foi apresentado um sistema de reconhecimento automático de comandos de fala independente de locutor e com vocabulário restrito. O sistema proposto baseia-se na extração de coeficientes mel-cepstrais da gravação correspondente à pronúncia e na classificação da mesma por meio de RNAs. Também foi proposta uma modificação no processo de detecção de palavras, onde alterou-se a forma de determinação dos limites do início e final da palavra nos dados de voz.

Os resultados dos experimentos realizados mostraram que a modificação proposta no algoritmo de detecção de palavras melhorou a taxa de classificação do sistema utilizado em relação ao algoritmo proposto na literatura. Evidencia-se que o novo cálculo dos limites da palavra consegue descartar mais eficientemente informações desnecessárias presente nas elocuções, além de descartar interferências geradas durante a produção das palavras, que é um problema apontado na literatura.

Para trabalhos futuros, sugere-se a avaliação do desempenho do algoritmo de detecção de palavras proposto em relação aos algoritmos apresentados em outros trabalhos e a implementação de algoritmos para redução de interferência nos sinais de áudio. Também sugere-se a avaliação de outros métodos além da RNA apresentada aqui para a etapa de classificação.

Agradecimentos

Os autores agradecem o apoio da Universidade do Estado de Santa Catarina e ao Laboratório de Aplicações Biomédicas e Sistemas Embarcados, onde foram feitas as gravações para a base de dados.

Referências

Bresolin, A. d. A. (2008). *Reconhecimento de voz através de unidades menores do que a palavra, utilizando Wavelet Packet e SVM, em uma nova estrutura hierárquica de decisão*, Tese de doutorado, Universidade Federal do Rio Grande do Norte.

Furui, S. (2005). 50 years of progress in speech and speaker recognition, *Proceedings of the 10th International Conference on Speech and*

Computer (SPECOM 2005), Patras, Grece, pp. 1–9.

Gold, B. and Morgan, N. (2000). *Speech and audio signal processing: processing and perception of speech and music*, John Wiley.

Haykin, S. (2001). *Redes neurais: princípios e prática*, 2ª edn, Bookman, Porto Alegre.

Igel, C. and Hüsken, M. (2000). Improving the rprop learning algorithm, *Proceedings of the Second International Symposium on Neural Computation*, p. 115–121.

Juang, B. and Rabiner, L. (2005). Automatic speech recognition: a brief history of the technology development, *Encyclopedia of Language and Linguistics, Elsevier* pp. 1–24.

Lamel, L. F., Rabiner, L. R., Rosenberg, A. E. and Wilpon, J. G. (1981). An Improved Endpoint Detector for Isolated Word Recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **29**(4): 777–785.

Lévy, C., Linarès, G. and Nocera, P. (2003). Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems, *IEEE Workshop on DSP in Mobile and vehicular systems*, Nagoya, Japan.

Nunes, H. F. (1996). *Reconhecimento de fala baseado em HMM*, Dissertação de mestrado, Universidade Estadual de Campinas.

Picone, J. (1993). Signal modeling techniques in speech recognition, *Proceedings of the IEEE* **81**(9): 1215–1247.

Rabiner, L. and Juang, B. (1993). *Fundamentals of Speech Recognition*, Prentice Hall Signal Processing Series, PTR Prentice Hall.

Rabiner, L. and Schafer, R. (1978). *Digital processing of speech signals*, Prentice-Hall signal processing series, Prentice-Hall.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm, *In Proceedings of the IEEE International Conference on Neural Networks*, pp. 586–591.

Rumelhart, D., Hinton, G. and Williams, R. (1985). *Learning Internal Representations by Error Propagation*, ICS report, Institute for Cognitive Science, University of California, San Diego.

Silva, I. N., Spatti, D. H. and Flauzino, R. A. (2010). *Redes Neurais Artificiais para Engenharia e Ciências Aplicadas - Curso Prático*, Artliber.

Tebelskis, J. (1995). *Speech recognition using neural networks*, Phd thesis, Carnegie Mellon University.

Zade, K. R. A., Ardil, C. and Rustamov, S. S. (2006). Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems, *Engineering and Technology* pp. 74–80.