UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO ENGENHARIA ELÉTRICA

CARLA MARIA MARTINS DOS SANTOS

Desenvolvimento de um sistema de reconhecimento de fala usando modelos ocultos de Markov

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO
2014

CARLA MARIA MARTINS DOS SANTOS

Desenvolvimento de um sistema de reconhecimento de fala usando modelos ocultos de Markov

Trabalho de conclusão de curso apresentado como requisito parcial à obtenção do título de engenheira elétrica da Universidade Tecnológica Federal do Paraná, Câmpus Cornélio Procópio.

Orientador: Prof. Dr. Paulo Rogério Scalassara.

CORNÉLIO PROCÓPIO
2014

Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio
Departamento de Engenharia Elétrica
Curso de Engenharia Elétrica

FOLHA DE APROVAÇÃO

Carla Maria Martins dos Santos

Desenvolvimento de um sistema de reconhecimento de fala usando modelos ocultos de Markov

Trabalho de conclusão de curso apresentado às 14:00hs do dia 26/11/2014 como requisito parcial para a obtenção do título de Engenheiro Eletricista no programa de Graduação em Engenharia Elétrica da Universidade Tecnológica Federal do Paraná. O candidato foi arguido pela Banca Avaliadora composta pelos professores abaixo assinados. Após deliberação, a Banca Avaliadora considerou o trabalho aprovado.

_____

Prof(a). Dr(a). Paulo Rogério Scalassara - Presidente (Orientador)

_____

Prof(a). Dr(a). Wagner Endo - (Membro)

_____

Prof(a). Dr(a). Fábio Renan Durand - (Membro)

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

Dedico este trabalho aos meus pais, Carlos e Edna; ao meu irmão, Guilherme e ao meu namorado e amigo, Alex.

Agradecimentos

A Deus e a Mãe Três Vezes Admirável de Shoenstatt, por me concederem vida, luz, saúde e força de vontade; por sempre escutarem minhas preces; pela proteção a mim e à minha família; e por serem a verdade no meu caminho.

Aos meus pais e meu irmão, pela base em todos os momentos, sem a qual eu não conseguiria ter chegado tão longe; pela compreensão e apoio incondicional em todos os meus projetos e sonhos; por nunca terem deixado de acreditar.

Ao meu namorado, pelo carinho, paciência, cuidado e força em todos os momentos.

Ao meu orientador, Profº. Dr. Paulo Rogério Scalassara, pelas lições compartilhadas, pelo companheirismo, bom humor e pela confiança a mim dispensada.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

HMM	*Hidden Markov Model*
MFCC	*Mel-Frequency Cepstral Coefficients*
GMM	*Gaussian Mixture Model*
ANN	*Artificial Neural Networks*
RAL	Reconhecimento Automático de Locutor
RAV	Reconhecimento Automático de Voz
PDF	*Probability Density Function*
FFT	*Fast Fourier Transform*
DTFT	*Discrete Time Fourier Transform*

Resumo

Neste trabalho, apresenta-se o desenvolvimento de um sistema de reconhecimento de fala no software Matlab capaz de reconhecer palavras pronunciadas de forma isolada por diferentes locutores. O método utilizado baseia-se em três etapas: pré-processamento dos sinais, modelagem por cadeias de Markov e reconhecimento de padrões. Porém, o sistema como um todo é formado por cinco blocos principais: aquisição do sinal, pré-processamento, extração dos parâmetros, Modelo Oculto de Markov (HMM) e reconhecimento do sinal de interesse, podendo evoluir para classificação da elocução/locutor e acionamento de algum dispositivo/sistema de interesse. Nos sistemas de reconhecimento de fala, os HMM são capazes de modelar as variabilidades do sinal de voz, baseando-se em um processo estocástico que pode realizar o reconhecimento de palavras isoladas ou contínuas, com vocabulários pequenos ou grandes. Os padrões usados foram os *Mel-Frequency Cepstral Coefficients* (MFCC) que utilizam coeficientes cepstrais de frequência mel para representar as características do sinal de voz. Derivados da Transformada Rápida de Fourier e da análise por meio de um banco de filtros na escala Mel, os MFCC são utilizados para treinar o HMM e validar o reconhecimento. Diante disso, para maior robustez do sistema foram coletadas amostras de diferentes usuários, formando-se um banco de dados mais completo para o treinamento e validação do HMM. Na análise dos resultados, a aplicação de HMM para treinamento e validação do sistema apresentou índice médio de acerto de 92% no reconhecimento da elocução de interesse, quando treinado com apenas um locutor e, índice de acerto de 98% quando treinado com todos os locutores usados na validação, o que demonstra que o sistema é capaz de realizar o reconhecimento independente do locutor.

Palavras-chave: Reconhecimento de fala; Modelos Ocultos de Markov; Reconhecimento de padrões; MFCC.

Abstract

In this study, we present the development of a speech recognition system in Matlab software that can recognize words spoken by different speakers. The method proposed is based on three stages: signal pre-processing, Markov chains and pattern recognition. However, the whole system consists of five main blocks: signal acquisition, preprocessing, parameter extraction, Hidden Markov Model and signal recognition of interest, it may evolve towards classification of speech/speaker and control some device/system of interest. In speech recognition systems, HMM is capable of modeling the variability of the speech signal, based on a stochastic process which may carry the recognition of isolated words or continuous, with small or large vocabularies. The patterns used were the Mel-Frequency Cepstral Coefficients using mel cepstral coefficients to represent characteristics of the speech signal. Derived from the Fast Fourier Transform and the analysis by means of a filter bank in mel scale, the MFCC are used to train the HMM and recognition validation. For greater robustness the system, samples were collected from different users, forming a database of more complete data for training and validation of the HMM. In analyzing the results, the application of HMM for training and validation of the system had a mean accuracy level of 92% in recognition of utterances of interest, when trained with only one speaker, and success rate of 98% when trained with all the speakers used in validation, which shows that the system is capable of speaker-independent recognition.

Key-words: Speech Recognition; Hidden Markov Model; Pattern recognition; MFCC.

1 Introdução

A fala é o modo natural de comunicação do ser humano, tendo sua forma dependente das estruturas fonológicas, sintáticas e prosódicas da língua, do ambiente acústico, do contexto em que é produzida e do canal pelo qual é propagada. Produzida de maneira diferente por cada pessoa, as variações da fala devidas ao dialeto, forma do trato vocal, ritmo de pronúncia, entre outros, fazem com que a implementação de sistemas capazes de reconhecê-la tenham a habilidade de lidar com todas as variabilidades que apresenta, já que o efeito causado por variáveis não modeladas ou mal modeladas pode ser considerado devastador (YNOGUTI, 1999).

Conforme Juang e Rabiner (2004), o projeto de uma máquina que imite o comportamento humano desperta a atenção de cientistas e engenheiros há muito tempo. Desde a década de 1930, quando Dudley (1939) e Dudley e Watkins (1939), dos laboratórios Bell, propuseram um modelo de sistema para análise e síntese de voz, o problema de reconhecimento automático de voz começou a progredir. Com os avanços na modelagem estatística nos anos 1980, os sistemas de reconhecimento de fala começaram a ter grande aplicação em tarefas que exigem uma interface homem-máquina.

As tentativas de desenvolvimento de máquinas para imitar a capacidade humana de comunicação iniciaram-se na segunda metade do século XVIII, sendo guiadas pela teoria da fonética acústica, com o objetivo de explicar a acústica de um discurso falado. Nessa época, o interesse principal era criar uma máquina que falasse, já que o que se tinha disponível eram tubos de ressonância acústica, usados para imitar o trato vocal (JUANG; RABINER, 2004).

Durante a primeira metade do século XX, Fletcher (1922) e cientistas dos Laboratórios Bell documentaram as características de um espectro de discurso, descrevendo suas peculariedades de acordo com a habilidade do ouvido humano. Com isso, em 1930, Fletcher desenvolveu um sintetizador de voz, considerado um importante marco na evolução das máquinas que falam. Por sua vez, Davis, Biddulph e Balashek (1952), também dos Laboratórios Bell, construíram um sistema de reconhecimento de dígito isolado para um único locutor, que realizava o reconhecimento com base na trajetória dos formantes de cada um dos dez dígitos (0 a 9). Nesse aspecto, cabe ressaltar, que a década de 1950 foi marcada pelo desenvolvimento de diversos

projetos de reconhecimento de fala ao redor do mundo, tais como os dos Laboratório RCA e Labotatórios NEC dos Estados Unidos, além do Laboratório de Rádio Pesquisa em Tóquio (JUANG; RABINER, 2004).

No final da década de 1960, Atal e Hanauer (1971) e Itakura e Saito (1970) formularam os conceitos da codificação por predição linear que de forma simplificada faz uma estimativa da resposta do aparelho vocal. A partir disso, em meados dos anos 1970, ideias básicas da aplicação dessa tecnologia de reconhecimento de padrões de voz começaram a ser propostos. Também nesse período, começou-se a utilização dos sistemas de reconhecimento em algumas empresas, fato que influenciou a ARPA (*Advanced Research Projects Agency*) dos Estados Unidos a financiar o primeiro programa de reconhecimento de fala, fazendo com que o desenvolvimento desses sistemas tivesse um crescimento significativo a partir de então (JUANG; RABINER, 2004).

Mesmo diante dessa significativa evolução, os sistemas de reconhecimento ainda enfrentavam dificuldades na descrição matemática do discurso, de forma que pesquisas paralelas eram desenvolvidas para se conseguir o domínio sobre as variabilidades acústicas que o sinal de voz apresenta. Com isso, começou-se a estudar métodos estatísticos, sendo que o desenvolvimento deles na década de 1980, principalmente da estrutura dos Modelos Ocultos de Markov (HMM), trouxeram grandes avanços na área. Assim, pode-se dizer que as pesquisas nos anos 1980 se caracterizaram por uma mudança de metodologia, sendo baseadas em um modelo maia rigoro para a modelagem estatística (JUANG; RABINER, 2004).

Nesses anos, embora já se tivesse a ideia dos HMM, sua metodologia ainda não estava completa, sendo definida apenas com a publicação da teoria alguns anos depois. Atualmente, conforme Juang e Rabiner (2004), a maioria dos sistemas de reconhecimento de voz se baseiam nessa estrutura estatística, a qual teve melhorias adicionais nos anos 1990 e continua sendo aprimorada com os avanços da tecnologia .

Os sistemas de reconhecimento de fala já possuem aplicações em diversas áreas, tendo-se a compreensão de que qualquer atividade de interação homem-máquina pode ser aprimorada por esses métodos. Conforme Silva (2009), eles já são comumente encontrados em dispositivos de controle e comando, de telefonia, de transcrição, além de estarem presentes em centrais de atendimento ao cliente e robótica, visando sempre facilitar a vida do homem. Contudo, a capacidade de reconhecimento desses sistemas ainda está abaixo da capacidade humana, principalmente quando se tem a presença de ruídos, distorções de canal, variações de pronúncia, entre outros, o que faz com que um dos maiores problemas do reconhecimento de fala seja o modelamento das características do sinal.

No entanto, mesmo com essas barreiras, a tecnologia do reconhecimento da informação tem invadido o cotidiano das pessoas. Como exemplo, tem-se o aplicativo SIRI da Apple que utiliza o reconhecimento de fala para realizar as funções de enviar mensagens SMS, escrever e-mails, realizar ligações, entre outras funções (SOTERO Jr., 2011). Outro aplicativo popular é o *Voice Search* da Google que permite a pesquisa por voz em sistemas computacionais. Além disso, tem-se também a assistente pessoal inteligente desenvolvida pela empresa Microsoft e denominada Cortana (MICROSOFT, 2014), que tem funções semelhantes ao SIRI.

Diante disso, neste trabalho apresenta-se o desenvolvimento e implementação de um sistema de reconhecimento de fala no software Matlab, que possua capacidade de processamento de sinais e possa ser aplicável em acionamento de outros sistemas de interesse. A extração de informações do sinal de voz é feita através da metodologia sugerida por Rabiner e Juang (1978) baseada na extração dos MFCC (*Mel-Frequency Cepstral Coefficients*) derivados da Transformada Rápida de Fourier e da análise por meio de um banco de filtros na escala Mel. Já para o reconhecimento, o sistema desenvolvido baseia-se no HMM que, na área do processamento de fala, são capazes de absorver as variações temporais das diferentes amostras da mesma palavra, garantindo maior robustez ao projeto.

1.1 Problema

A busca por sistemas de comunicação entre os seres humanos e entre ser humano e máquina é, conforme Rabiner e Schafer (2007), algo que desperta a atenção do homem desde antes da invenção do telefone por Alexander Graham Bell.

Com a contribuição de novos métodos de processamento digital de sinais, a partir de 1960, grandes avanços passaram a ocorrer no processamento digital de voz, sendo que este se inicia com a análise da natureza básica do sinal chegando às aplicações em comunicação de voz e síntese e reconhecimento automático da expressão (RABINER; SCHAFER, 2007).

Nos sistemas de comunicação, a informação é codificada em uma forma de onda que pode ser transmitida, gravada, manipulada e também decodificada para um ouvinte humano. No processo de fala, por sua vez, a forma analógica fundamental da mensagem é uma forma de onda acústica, chamada sinal de voz, sendo que, a partir desse sinal é possível processar a fala (RABINER; SCHAFER, 2007).

Conforme Silva (2009), o reconhecimento automático de fala se traduz na conversão de um sinal acústico produzido pela fala humana em um sinal digital de áudio, por meio de um hardware, associado a um software, que identificará o conjunto de palavras faladas. A partir

disso, o problema de reconhecimento de fala envolve a coleta de áudio, o reconhecimento e, por fim, a aplicação. Na segunda etapa, reconhecimento de fala, tem-se a maior dificuldade já que a voz humana apresenta natureza interdisciplinar, fazendo com que a variabilidade dos sinais de fala seja um limitador de desempenho dos sistemas de reconhecimento.

Essa limitação de desempenho, conforme Silva (2009), relaciona-se principalmente a:

- variabilidade dos sons para um único locutor e entre locutores diferentes;
- variabilidade do transdutor e do canal, como microfones, telefones fixos e celulares;
- variabilidade do ruído de fundo gerado a partir de outras vozes, carros, ar-condicionado, dentre outros;
- variabilidade na produção da fala incluindo barulhos resultantes de movimentos da boca, ruídos de respiração, hesitações ao falar, etc.

A partir do exposto, o problema a ser tratado neste trabalho consiste em analisar corretamente o sinal de voz de interesse, abrangendo todas as etapas de aquisição e processamento do sinal, a fim de que se possa aplicá-lo na execução da tarefa pretendida pelo locutor.

1.2 Justificativa

O reconhecimento de fala surgiu como um atrativo para sistemas computadorizados, tendo sua história relacionada com a evolução dos microcomputadores e processamento digital de sinal (TAFNER, 1996).

A grande evolução na área de reconhecimento de voz começou com o desenvolvimento de novas técnicas de processamento digital de sinais, disponibilidade de computadores rápidos e mais baratos, desenvolvimento de padrões para avaliação de desempenho, além da maturidade alcançada em algumas técnicas como HMM, Modelos de Mistura de Gaussiana (GMM) e Redes Neurais Artificiais (ANN) (SILVA, 2009).

Rabiner e Schafer (2007) expõem que uma grande classe de aplicação do processamento digital de voz é dedicada à extração de informações decorrentes de fala. Por isso, muitos sistemas reconhecem a fala com o objetivo de extrair a mensagem ou identificar quem está falando. Isso faz com que as técnicas desse processamento digital envolvam o armazenamento e transmissão digital, síntese de fala, reconhecimento de fala, identificação e verificação do

falante, melhoria da qualidade de fala e, por fim, ajuda aos deficientes. Nesse último caso, pode-se citar os deficientes visuais que são capazes de utilizar um computador que lhes apresente as informações de forma audível, permitindo-lhes também criar mensagens no editor de texto pelo simples uso de sua fala.

Diante disso, pode-se dizer que os sistemas de reconhecimento de fala já possuem sua importância e espaço garantidos no mundo atual, devido a melhoria que podem trazer para a vida do ser humano. Assim, o desenvolvimento de sistemas capazes de operar com alto desempenho em qualquer tarefa e/ou ambiente acústico é o que desafia o estado da arte da tecnologia atual e motiva o estudo e desenvolvimento dessa área.

1.3 Objetivos

O objetivo deste trabalho é desenvolver um sistema de reconhecimento de fala independente de locutor baseado nos coeficientes mel-cepstrais e Modelos Ocultos de Markov no software Matlab, que possua capacidade de processamento de sinais e possa ser aplicável em acionamento de outros sistemas de interesse.

Os objetivos específicos são:

- Processar o sinal de voz para extração dos coeficientes mel-cepstrais.
- Realizar o treinamentos dos Modelos Ocultos de Markov.
- Validar o reconhecimento de fala através de testes com diferentes locutores.
- Implementar o sistema para o acionamento de outros sistemas.

1.4 Estrutura do trabalho

Para alcançar o proposto, tem-se a divisão do trabalho em seis capítulos.

No Capítulo 2, discorre-se sobre a teoria do reconhecimento de fala, apresentando suas características e classificações, bem como alguns trabalhos já desenvolvidos na área. Além disso, tem-se a teoria dos HMM, com sua estrutura e exemplo.

No Capítulo 3, tem-se as etapas necessárias para o desenvolvimento do sistema de reconhecimento de fala, explicando e demonstrando a influência de cada etapa no sinal de voz.

Na sequência, tem-se no capítulo 4 os resultados do sistema de reconhecimento de fala, com os experimentos e discussões.

Por fim, no capítulo 5, apresenta-se a conclusão do trabalho com as sugestões para novas pesquisas na área.

2 Teoria

Neste capítulo apresenta-se conceitos importantes da área do reconhecimento de fala, que consiste no processo de conversão de um sinal acústico produzido pelo homem em um sinal digital de áudio através de um hardware associado a um software, o qual a partir de uma base de dados identificará a palavra falada. Depedendo da aplicação, a palavra reconhecida pode ser o resultado final do sistema ou, no caso de projetos de comando, a entrada de outros sistemas (SILVA, 2009).

2.1 Classificação do Reconhecimento de Fala

Conforme Valiati (2000), o estudo da fala pode ser dividido em três áreas principais: análise, síntese e reconhecimento, sendo que este último subdivide-se em reconhecimento automático de locutor (RAL) e reconhecimento automático de voz (RAV).

O RAL realiza o reconhecimento de indivíduos por meio da verificação de elocuções ou pela extração das características distintas de cada locutor. Desta maneira, esses sistemas podem ser aplicados ao controle de acesso a determinadas áreas restritas.

Por sua vez, o RAV caracteriza-se pela compreensão de uma elocução, destinando-se a sistemas de informações, tais como os que necessitam de resposta a tipos de questões sim/não. Sua classificação também pode referir-se à dependência ou não do locutor.

Outra maneira de se classificar os sistemas de reconhecimento de fala, conforme (SILVA, 2009) é de acordo com seu objetivo:

- Identificação de palavras: o sistema busca reconhecer a mesma palavra, independentemente do locutor, podendo ser aplicado em sistemas que dependem de comandos vocais de diferentes usuários.
- Identificação de pessoas: o sistema é capaz de reconhecer a pessoa por meio da palavra pronunciada, sendo importante para sistemas de segurança, tal como os que envolvem o

controle de acesso a determinados ambientes.

- Identificação de pessoa/palavra: o sistema reconhece determinada palavra apenas quando pronunciada por determinada pessoa. Esse modelo é considerado como o mais simples podendo ser aplicado em diversos sistemas.

Também de acordo com Silva (2009), dentre as classificações mais importantes tem-se as referentes ao estilo de pronúncia que o sistema aceita, ao tamanho do vocabulário e à dependência ou independência do locutor, conforme detalhado a seguir:

1. Quanto ao modo de pronúncia:
    - Reconhecedor de palavras isoladas: o sistema reconhece apenas uma palavra por vez ou palavras onde se tenha uma pausa mínima entre elas, sendo, por isso, o mais simples de ser implementado.
    - Reconhecedor de palavras conectadas: o sistema utiliza palavras como unidade fonética padrão, sendo capaz de reconhecer sentenças completas pronunciadas sem pausa entre as palavras.
    - Reconhecedor de fala contínua: o sistema é capaz de reconhecer a comunicação natural, devendo ser capaz de lidar com todas as características e vícios da linguagem natural.
2. Quanto ao tamanho do vocabulário:
    - Vocabulário pequeno: reconhecimento de até 20 palavras.
    - Vocabulário médio: reconhecimento de 20 até 100 palavras.
    - Vocabulário grande: reconhecimento de 100 até 1000 palavras.
    - Vocabulário muito grande: reconhecimento de mais de 1000 palavras.
3. Quanto à dependência de locutor:
    - Dependente de locutor: o sistema é capaz de reconhecer a fala somente de pessoas cujas vozes foram utilizadas em seu treinamento.
    - Independente de locutor: há o reconhecimento da palavra pronunciada por qualquer pessoa com uma taxa de acerto aceitável. Esse sistema deve ser treinado com dados de pessoas com diferentes idades, sexo, sotaques, etc.

No presente trabalho, desenvolveu-se um sistema de identificação de palavras que reconhece uma elocução pré-definida, podendo ser aplicado à realização de alguma tarefa de interesse.

2.2 Aplicações do Reconhecimento de Fala

Em Tevah (1996), afirma-se que existem diversas aplicações que utilizam ou poderiam utilizar sistemas de reconhecimento de voz, destacando como exemplos comuns a transcrição de texto, comandos de dispositivos por voz, atendimento eletrônico por telefone, biometria, entre outros. Diante disso, tem-se a distribuição dessas aplicações em dois sistemas: sistemas de interface e sistemas transcritores.

Os sistemas de interface utilizam as técnicas de reconhecimento de fala para ações de comando ou navegação que facilitam o acesso de usuários convencionais e viabilizam o uso de computadores por deficientes físicos (TEVAH, 1996).

Atualmente, esses sistemas fazem parte do dia-a-dia das populações, estando presentes em celulares que discam através de comandos de voz e centrais telefônicas que processam informações ditas por seus usuários.

Por sua vez, os sistemas transcritores realizam as funções de captação e transcrição em linguagem corrente do que está sendo falado pelo usuário, não tendo o objetivo, portanto, de realizar determinada ação. O funcionamento desse sistema é comparado ao de uma secretária que apenas toma nota do que está sendo dito (TEVAH, 1996).

A evolução desses sistemas mostra-se importante para o desenvolvimento de um sistema de conversação telefônica para deficientes auditivos, sendo composto por um módulo de síntese e reconhecimento de voz que pode facilitar a integração deles à sociedade.

Em Tafner (1996), propõe-se um sistema de reconhecimento de palavras isoladas pronunciadas por determinado locutor. Nesse experimento, realizou-se a extração de dados das amostras para redução da quantidade de informações, conseguindo-se eliminar o ciclo negativo do sinal amostrado, detectar a forma de onda, medir o sinal amostrado e normalizar o sinal medido. A partir da efetivação dessas etapas, os dados foram treinados e validados na rede neural Kohonen, tendo índices de acerto na faixa de 80% a 90%.

Em Silva (2009), trabalho utilizado como base, apresenta-se um sistema de reconhecimento de palavras isoladas baseado em HMM e nos coeficientes mel-cepstrais, testado para um vocabulário pequeno formado pelos dígitos de 0 a 9 e pelas palavras “sim” e “não”. Na apresentação dos resultados, o autor utiliza a técnica de validação *K-fold cross-validation* e um limiar para a máxima verossimilhança, que torna o sistema mais confiável. Além disso, o trabalho utiliza o HMM com observações contínuas, modeladas por uma quantidade finita de misturas de gaussianas.

2.3 Modelos Ocultos de Markov

2.3.1 Introdução

Baum e Petrie (1966), Baum e Eagon (1967) e Baum et al. (1970), no final da década de 1960, publicaram artigos lançando as bases para o formalismo dos Modelos Ocultos de Markov. Nas primeiras aplicações dessa modelagem, visou-se o reconhecimento de fala sendo que os trabalhos de Jelinek (1976) e Baker e Bahl. (1975) foram os pioneiros no uso de HMM. A partir da década de 1980, o HMM passou a ser aplicado no sequenciamento de DNA, alcançando assim grande importância no campo da bioinformática (ESPINDOLA, 2009).

Baseado em modelamento estocástico, os HMM caracterizam-se pela adaptação às variabilidades do sinal de fala e pela flexibilidade em modelar palavras. Assim, são capazes de realizar o reconhecimento de palavras isoladas ou contínuas, com vocabulários pequenos ou grandes. O HMM tem também a capacidade de assimilar informações de uma grande quantidade de dados de treinamento, conseguindo, portanto, lidar com diversas elocuções de referência simultaneamente (YOMA, 1993).

Conforme Espindola (2009), os HMM são um tipo especial de processos estocásticos com aplicabilidade diversa, podendo ser encontrados em aplicações químicas, biológicas, físicas, entre outras.

Em um sistema estatístico de reconhecimento de fala, geralmente as palavras são representadas por um conjunto de modelos probabilísticos de unidades linguísticas, tal como os fonemas (YNOGUTI, 1999). A partir de uma sequência de padrões acústicos, extraídos da elocução, pode-se fazer a concatenação de processos elementares, baseado nos HMM.

Dessa forma, um HMM é uma composição de dois processos estocásticos: uma cadeia de Markov oculta, relacionada à variação temporal e um processo observável, relacionado à variabilidade espectral (YNOGUTI, 1999).

2.3.2 Estrutura de um HMM

Um HMM é definido como um par de processos estocásticos *(X,Y)*, no qual o processo *X* é uma cadeia de Markov de primeira ordem, não diretamente observável (oculta), e o processo *Y* é uma sequência de variáveis aleatórias que assumem valores no espaço de parâmetros acústicos (observações) (YNOGUTI, 1999).

A partir de suas sequências de observações definidas, o HMM salta de um estado para o

outro, emitindo uma observação em cada estado. No reconhecimento de fala, o modelo pode ser visto como um espaço de estados finito no qual cada estado representa uma palavra, um fonema ou parte de um fonema, havendo troca de estado a cada unidade de tempo *t* (CARVALHO; SANTOS, 2014).

As saídas geradas por cada estado são vetores acústicos que possuem uma função de densidade de probabilidade associada, e são caracterizadas por um GMM (CARVALHO; SANTOS, 2014).

Como exemplo, na Figura 2.1 tem-se um modelo de HMM de 3 estados, denominado de *left-right*, que como o nome propõe, caminha sempre da esquerda para direita. Em geral,para o reconhecimento de fala utiliza-se esse modelo (YNOGUTI, 1999), também conhecido por modelo de Bakis.

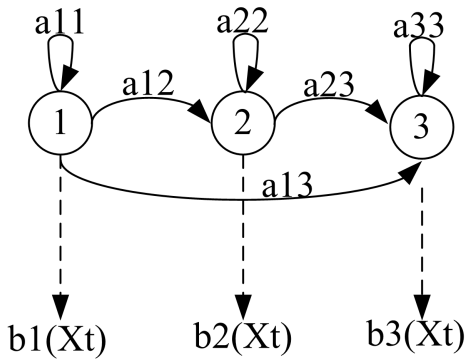

Figura 2.1: Modelo *left-right* para HMM.
Fonte: Adaptado de Carvalho e Santos (2014)

Na utilização do HMM tem-se duas formas ligeiramente diferentes para a emissão de uma observação. A primeira, utlizada no processamento acústico ou modelamento de sinais, emite uma observação no instante de chegada ao estado, denominada de máquina de Moore. A outra forma, utilizada em processamento de linguagem, emite uma observação durante a transição, sendo chamada de máquina de Mealy (YNOGUTI, 1999).

Seguindo a tendência geral dos sistemas de reconhecimento de fala (YNOGUTI, 1999), utiliza-se neste trabalho a forma de Moore.

No desenvolvimento desse processo, duas simplificações podem ser adotadas para a teoria do HMM, formuladas da seguinte maneira (DELLER Jr.; HANSEN; PROAKIS, 1993):

- Hipótese de Markov de primeira ordem: a história não tem influência na evolução futura da cadeia se o presente é especificado.
- Hipótese de independência das saídas: nem a evolução da cadeia nem as observações

passadas influenciam a observação atual se a última transição da cadeia é especificada.

Em ambas essas hipóteses, o HMM pode ser escrito da seguinte maneira (YNOGUTI, 1999): seja $y \in Y$ (variável que representa as observações) e $i, j \in X$ (variáveis que representam os estados do modelo). Então, pode-se representar o HMM por:

$$\mathrm{A} = \left\{ a_{ij} \mid i, j \in X \right\}$$

$$\mathrm{B} = \{ b_i(y) \mid i \in X, y \in X \}$$

$$\prod = \{ \pi_i \mid i \in X \}$$

onde A é a matriz de probabilidades de transição de estados (que carrega os coeficientes *a12, a22, a23* (CARVALHO; SANTOS, 2014)), B é a matriz de densidades de probabilidade de emissão de símbolos b e $\prod$ é a matriz de probabilidades de um modelo ser iniciado a partir de determinado estado, ambas definidas pelas Equações (2.1), (2.2) e (2.3):

$$a_{ij} \equiv P(X_t = j \mid X_{t-1} = i) \tag{2.1}$$

$$b_j(y) \equiv p(Y_t = y \mid X_t = j) \tag{2.2}$$

$$\pi_i \equiv P(X_0 = i) \tag{2.3}$$

Além disso, tem-se também os parâmetros μ e Σ das gaussianas associados a cada vetor emitido, os quais representam, respectivamente, as médias e variâncias do GMM (CARVALHO; SANTOS, 2014).

O HMM também pode ser classificado como discreto, semi-contínuo ou contínuo, dependendo do tipo de distribuição associada às probabilidades de emissão de símbolos (SILVA, 2009).

- Discreto: as observações são discretas por natureza ou discreticadas por quantização vetorial, gerando assim *codebooks*.
- Contínuo: as observações são contínuas, com função de densidade de probiblidade (PDF) também contínua, usualmente modelada como uma mistura finita de M gaussianas multidimensionais.

- Semi-contínuo: modelo intermediário entre o contínuo e o discreto.

Pelo fato de a utilização de HMM com densidades contínuas apresentarem melhores resultados para a área de reconhecimento de voz (SILVA, 2009), adotou-se esse modelo no presente trabalho .

2.3.3 Exemplo de um HMM

Em Rabiner (2002) tem-se um exemplo simples para ilustrar a aplicação de Cadeias de Markov. O exemplo trata da modelagem do tempo no decorrer dos dias.

Assim, seja uma variável estocástica X representando o clima e tendo suas possibilidades definidas em um conjunto discreto dado por: $\{S_1$ = chuvoso, S_2 = nublado, S_3 = ensolarado$\}$. Considerando-se que as observações são feitas um vez ao dia, que o resultado obtido será sempre um único desses três estados possíveis, sem combinação entre estados, e que as probabilidades de transição entre esses estados são dadas pela matriz:

$$\mathrm{A} = a_{ij} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

Adotando-se que o dia 1 é ensolarado, pode-se perguntar: qual a probabilidade de que o tempo para os 7 dias seguintes seja ensolarado-ensolarado-chuvoso-chuvoso-ensolarado-nublado-ensolarado?

Para resolução da questão pode-se definir a sequência de observação O = $\{X_0 = S_3, X_1 = S_3, X_2 = S_3, X_3 = S_1, X_4 = S_1, X_5 = S_3, X_6 = S_2, X_7 = S_3\}$. Após isso, o que se deseja, então, é obter a probabilidade de O, tendo-se o modelo:

$$\begin{aligned} P(O\|Modelo) &= P(S_3,S_3,S_3,S_1,S_1,S_3,S_2,S_3\|Modelo) \\ &= P(S_3)P(S_3\|S_3)P(S_3\|S_3)P(S_1\|S_3)P(S_1\|S_1)P(S_3\|S_1)P(S_2\|S_3)P(S_3\|S_2) \\ &= \pi_3.a_{33}.a_{33}.a_{31}.a_{11}.a_{13}.a_{32}.a_{23} \\ &= 1.(0.8)(0.8)(0.1)(0.4)(0.3)(0.1)(0.2) \\ &= 1.536 \times 10^{-4} \end{aligned}$$

onde π_i = P($X_0 = S_i$), 1 $\leq$ i $\leq$ N indica a probabilidade inicial de cada estado.

Como se nota, no exemplo dado é permitido ao observador verificar a condição do

tempo de determinado dia, obtendo um dos estados da Markov como resposta. Por sua vez, nos HMM a evolução da cadeia fica escondida do observador, que não pode observar as condições do modelo.

Com o exposto, no próximo capitulo, tem-se a definição das etapas necessárias para o desenvolvimento do sistema de reconhecimento de fala.

3 Materiais e métodos

Neste capítulo, apresentam-se as etapas necessárias para o desenvolvimento de um sistema de reconhecimento de fala, iniciando com a aquisição do sinal até a extração dos parâmetros necessários para treinamento e validação do sistema.

3.1 Processamento da Fala

Para o desenvolvimento de um sistema de reconhecimento eficiente é preciso que o sinal de voz seja corretamente processado. Conforme a Figura 3.1, o reconhecimento de fala pode ser dividido em quatro etapas: aquisição do sinal de voz, pré-processamento, extração de padrões e reconhecimento do sinal de fala.

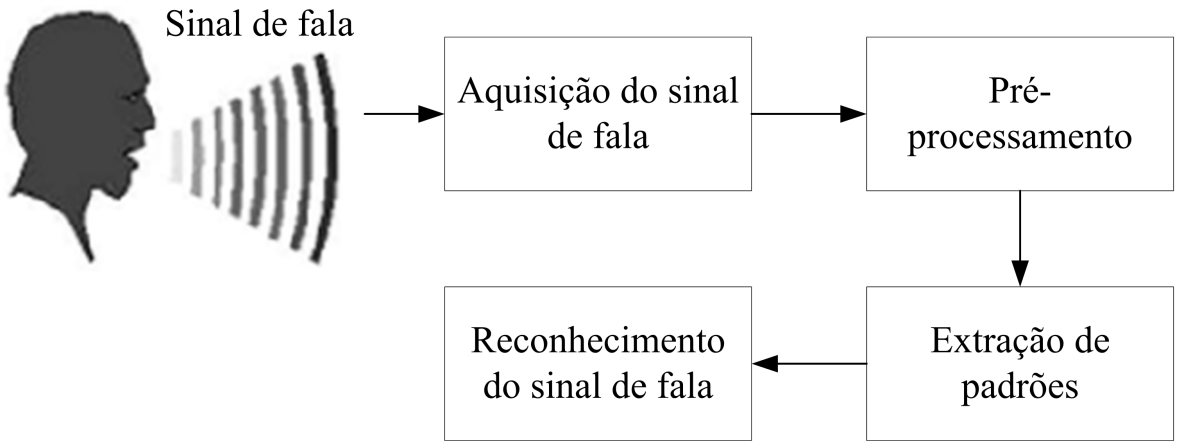

Figura 3.1: Sistema de Reconhecimento de fala.
Fonte: Adaptado de Zanotelli (2008).

3.1.1 Aquisição do Sinal de Voz

A primeira etapa, aquisição do sinal de voz, é realizada por meio de um transdutor, comumente um microfone, que faz a conversão dos sinais sonoros em sinais elétricos. Esses

sinais elétricos, por sua vez, são filtrados por um filtro *anti-aliasing* que diminui os componentes de frequência superiores à metade da frequência de amostragem, permitindo a conversão do sinal para digital. Após a conversão do sinal, o mesmo pode ser então processado.

Para a gravação dos sinais adotou-se 22050 Hz para a frequência de amostragem e formato wave, através da função *wavrecord* do Matlab. Como exemplo de um dos sinais adquiridos, tem-se o apresentado na Figura 3.2.

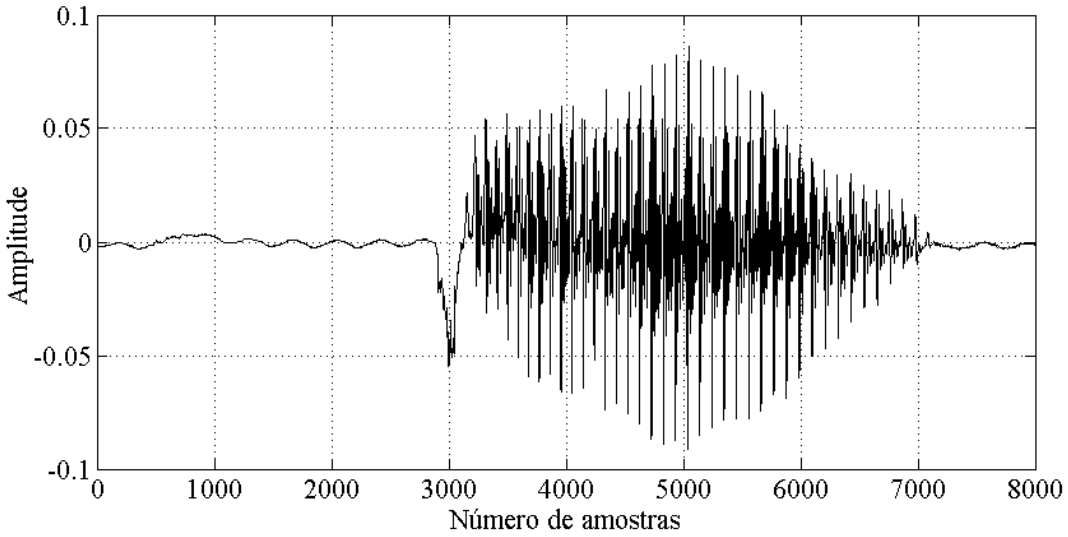

Figura 3.2: Sinal de voz captado pelo sistema.
Fonte: Autoria própria.

A partir da aquisição dos sinais, formou-se o banco de dados do sistema de reconhecimento sendo que, neste presente trabalho, utilizaram-se 350 elocuções.

3.1.2 Pré-processamento

Na etapa de aquisição do sinal de voz os dados necessários para o sistema de reconhecimento sofrem influência do ambiente de gravação e do canal de comunicação. Diante disso, faz-se necessário um pré-processamento do sinal, formado pelas etapas da Figura 3.3, com o intuito de deixá-lo mais próximo da fala pura.

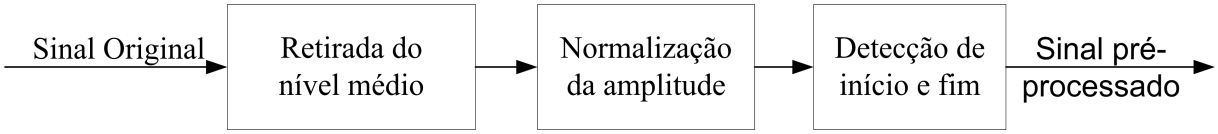

Figura 3.3: Diagrama de blocos do pré-processamento.
Fonte: Adaptado de Silva (2009).

A retirada do nível médio é importante porque o sinal de fala, em sua grande parte,

apresenta um componente contínuo que interfere na comparação entre os valores do sinal. Assim, para que todos os valores oscilem em torno de zero, faz-se sua retirada calculando-se a média aritmética das amplitudes do sinal e, depois, subtraindo de cada amplitude esta média.

Na subetapa de normalização da amplitude, se realiza a divisão de cada amostra do sinal pelo seu maior valor absoluto, ajustando o sinal entre -1 e 1. Dessa forma, se garante que todos os sinais analisados tenham a mesma faixa de intensidade sonora.

Por fim, o pré-processamento encerra-se com a detecção do início e fim da palavra a ser reconhecida. Essa subetapa é importante pois, ao se remover os períodos de silêncio existentes no início e no final do sinal, se conseque diminuir o tempo computacional gasto no reconhecimento, além de se eliminar possíveis ruídos ou sinais indesejados que possam existir. No presente trabalho, a detecção do início e fim do sinal foi feita de forma manual, com todos os sinais possuindo 11.025 amostras.

3.1.3 Extração de Padrões

O projeto de qualquer sistema de reconhecimento de fala possui a etapa de extração de padrões que se destina a obter apenas as informações realmente necessárias para a caracterização do sinal de voz. Isso se faz necessário, pois o sinal possui uma grande quantidade de dados que podem ser redundantes ou sem nenhuma significância para a distinção fonética. A característica fundamental desta etapa, conforme Silva (2009), é representar as unidades de fala com o menor número possível de padrões que contenham informações suficientes para caracterizar o sinal.

A extração dos padrões pode ser feita por técnicas de análise espectral, tais como a Transformada Rápida de Fourier (*Fast Fourier Transform* ou FFT), métodos de bancos de filtro (*Filter Banks)*, análise homomórfica ou análise cepstral (*mel-cepstrum)* e codificação por predição linear (*Linear Predictive Coding* ou LPC) (RABINER; JUANG, 1993), (RABINER; SCHAFER, 1978).

No entanto, dentre essas técnicas, a que é capaz de oferecer uma melhor metodologia para a separação do sinal de excitação da resposta impulsiva do trato vocal é a análise cepstral, caracterizada pelos MFCC, a qual atualmente é a técnica mais popular na área do reconhecimento de fala (BOUROUBA E-H., 2006) e, por isso, foi adotada nesse trabalho.

Para a obtenção dos parâmetros MFCC é necessário realizar as etapas ilustradas na Figura 3.4, descritas a seguir.

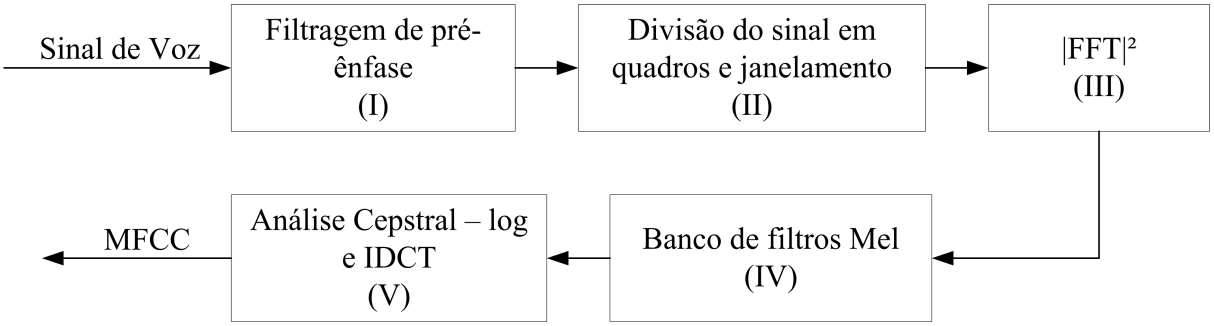

Figura 3.4: Diagrama de blocos para extração de padrões.
Fonte: Adaptado de Silva (2009).

I - Filtragem de Pré-ênfase

A filtragem de pré-ênfase tem o objetivo de atenuar as altas frequências por meio de um filtro passa-altas de primeira ordem representado pela função de transferência da Equação (3.1).

$$H(z) = 1 - \alpha z^{-1} \tag{3.1}$$

Com o valor de α igual a 0,95 tem-se a resposta em frequência do filtro na Figura 3.5. O efeito de pré-ênfase pode ser observado na Figura 3.6, bem como com a análise dos respectivos espectros de frequência (Figura 3.7).

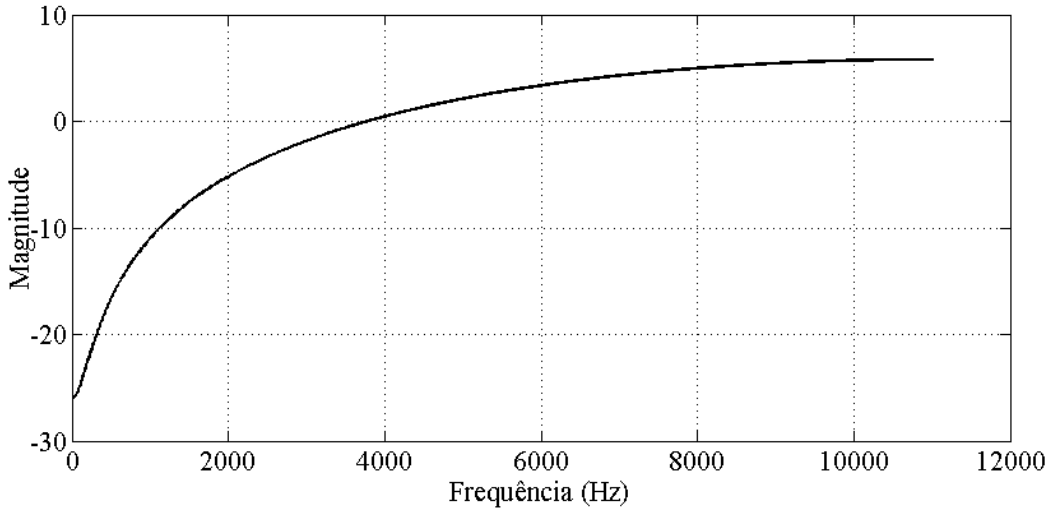

Figura 3.5: Resposta de frequência do filtro de pré-ênfase.
Fonte: Autoria própria.

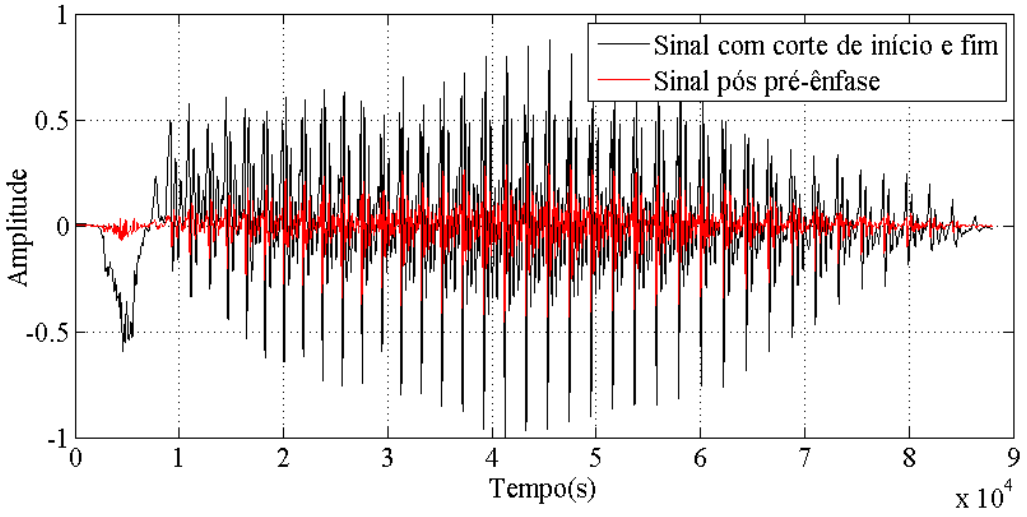

Figura 3.6: Comparação do sinal original e após aplicação do filtro de pré-ênfase.
Fonte: Autoria própria.

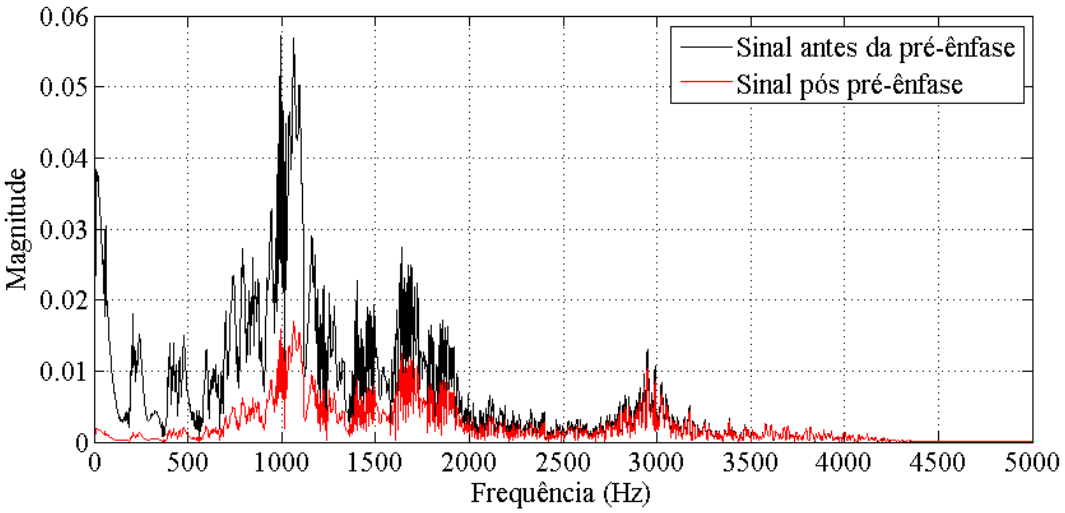

Figura 3.7: Comparação do espectro de frequência do sinal original e após a aplicação do filtro de pré-ênfase.

Fonte: Autoria própria.

II - Divisão do Sinal em Frames e Janelamento

O sinal de voz tem características variantes no tempo, enquanto que a extração de padrões apresenta bons resultados para sinais estacionários. Diante disso, é necessário dividir o sinal de entrada em quadros (*frames*) de 10 a 25 ms de forma que o sinal possa ser considerado quase estacionário nesse intervalo (BRAGA, 2006).

Essa divisão do sinal em quadros é feita pelo método de janelamento, que consiste em multiplicar o sinal pela função da janela utilizada (RABINER; JUANG, 1993). Neste trabalho,

usa-se a janela de Hamming (VALIATI, 2000), conforme a Figura 3.8, com 256 amostras, se obtendo quadros de aproximadamente 11,6 ms.

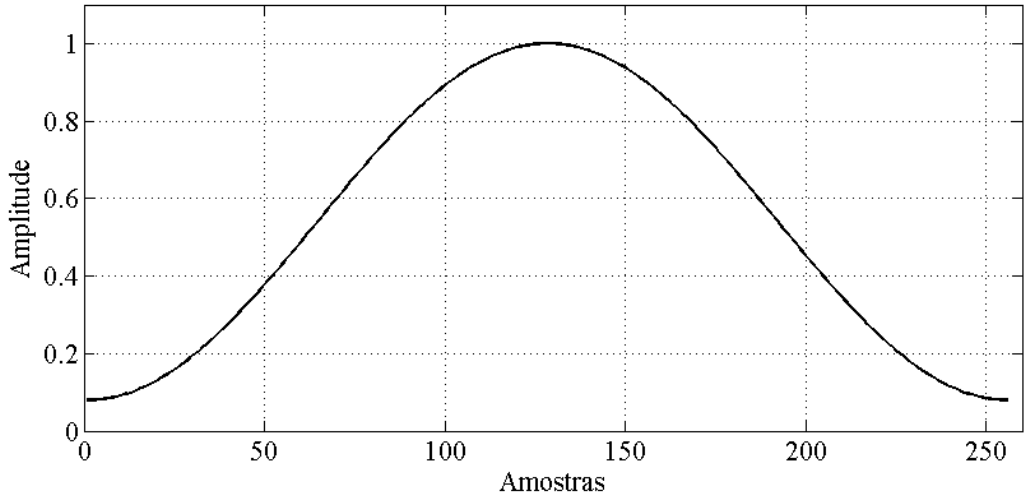

Figura 3.8: Janela de Hamming.
Fonte: Autoria própria.

Analisando-se o quadro 8 do sinal de voz, tem-se a Figura 3.9 na qual pode-se observar a atenuação nas extremidades do quadro.

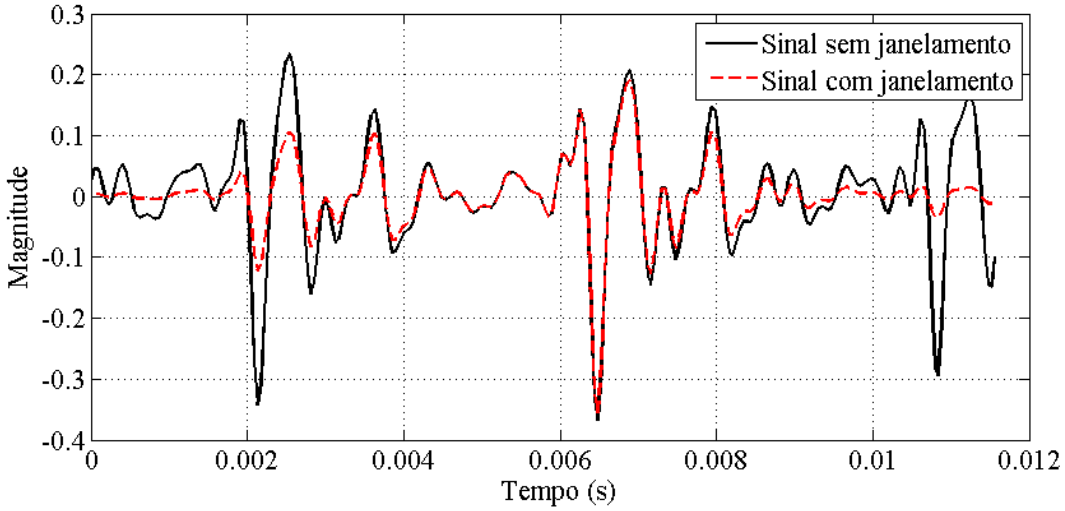

Figura 3.9: Sinal de voz após a Janela de Hamming.
Fonte: Autoria própria.

III - Transformada Rápida de Fourier

A Transformada Rápida de Fourier (FFT) é um algoritmo computacional que realiza a Transformada de Fourier de Tempo Discreto (*Discrete Time Fourier Transform* - DTFT), mapeando o sinal para o dompinio da frequência (HAYKIN; VEEN, 2001).

Ao sinal no domínio da frequência aplica-se o operador módulo, descartando-se a informação da fase, que pode ser desprezada para trabalhos de reconhecimento de fala (ZANOTELLI, 2008). E, por fim, se aplica o operador potência de 2, que equivale ao espectro de frequência do sinal, obtendo-se a análise da Figura 3.10.

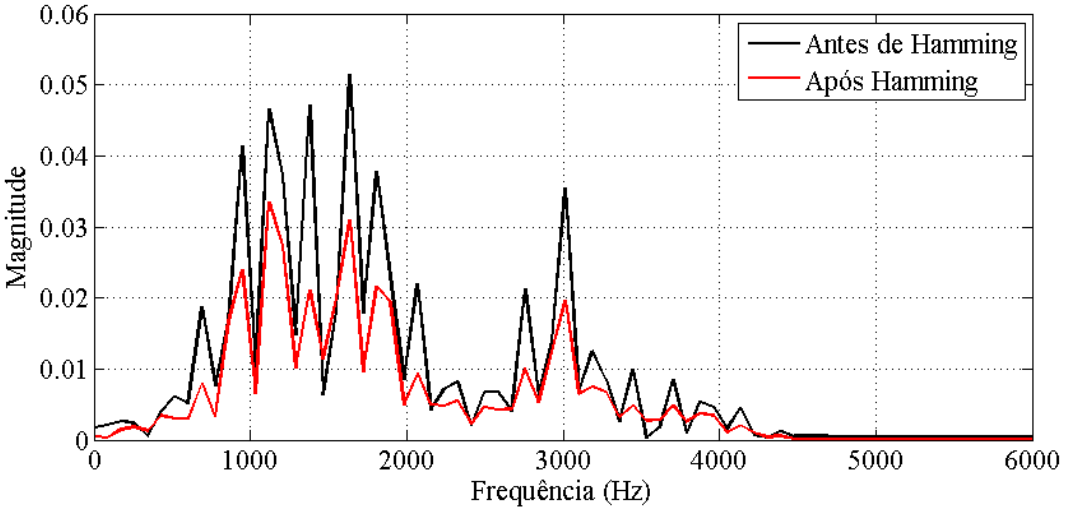

Figura 3.10: Espectro do sinal de voz após a Janela de Hamming - Quadro 8.
Fonte: Autoria própria.

IV - Bancos de Filtros na Frequência Mel

A técnica dos padrões mel-cepstrais baseia-se na modificação do espectro de voz conforme a escala Mel pelo fato da percepção para frequências sonoras não seguir uma escala linear tal como a escala Hertz (Hz) (PICONE, 1993). Assim, a escala Mel corresponde às frequências em Hz representadas em um valor medido em mel, que é sua unidade de frequência. A conversão para a escala Mel é feita através da Equação (3.2), na qual f é a frequência em Hz e *mel(f)* o valor em mel. A escala mel é considerada como linear de 0 a 1000 Hz, e logarítmica além de 1000 Hz (OLIVEIRA, 2001).

$$\text{mel(f)} = 1127 \ln\left(1 + \frac{f}{700}\right) \tag{3.2}$$

A obtenção dos parâmetros MFFC é feita por meio do cálculo do quadrado do módulo da FFT das amostras de cada quadro em análise, como já explicado. Depois, cada quadro é filtrado por um banco de filtros triangulares na escala Mel.

No presente trabalho as amostras foram filtradas por 39 filtros triangulares, conforme a Figura 3.11.

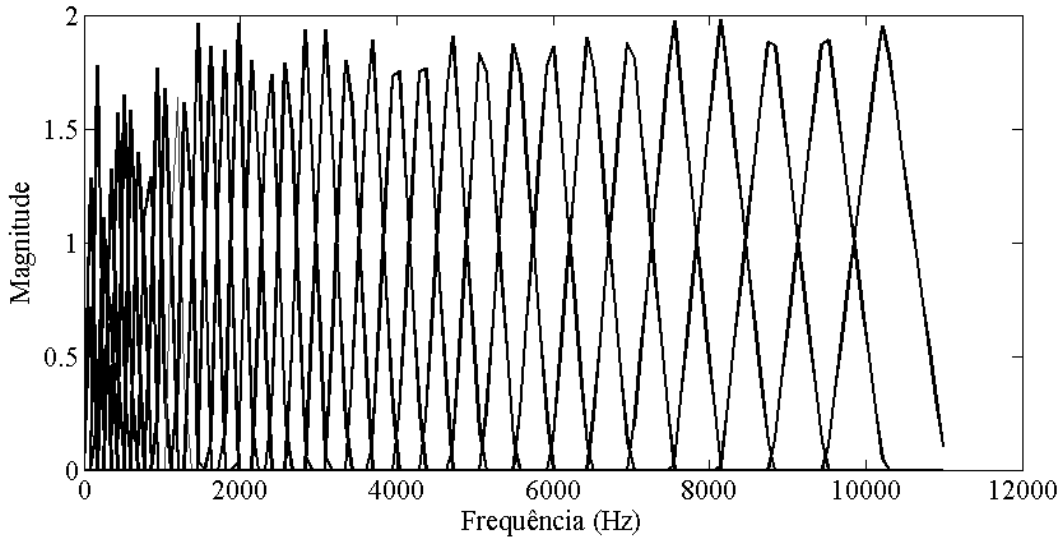

Figura 3.11: Banco de 39 filtros triangulares na escala Mel.
Fonte: Autoria própria.

V - Análise Cepstral

A análise cepstral é, conforme Oppenheim e Schafer (2012), uma técnica não linear que tem-se provado útil em diversas aplicações, oferecendo flexibilidade e sofisticação consideráveis na tecnologia de processamento em tempo discreto de sinais.

Em Bogert, Healy e Tukey (1963) foi observado que o logaritmo do espectro de potência de um sinal contendo um eco possui um componente periódico aditivo devido ao eco e, portanto, o espectro de potência do logaritmo do espectro de potência deve exibir um eco. Há essa função Bogert, Healy e Tukey (1963) deram o nome de *cepstrum*, trocando as letras da palavra *spectrum* (espectro), pois, "em geral, operamos no lado da frequência de maneiras comuns no lado do tempo e vice-versa".

Desta forma, desde a introdução do cepstrum, seus conceitos provaram-se úteis na análise de sinais, sendo aplicados com sucesso no processamento de sinais de voz (OPPENHEIM; SCHAFER, 2012).

Para obtenção dos cepstros, se calcula o logaritmo da magnitude na saída dos filtros, acumulando-se esse valor. (SILVA, 2009). Como no trabalho serão utilizados os MFCC, é necessário, após a obtenção dos cepstros, aplicar a Transformada Cosseno Discreta Inversa (IDCT) para produzir o vetor de característica formado pelos coeficientes mel-cepstrais (ZANOTELLI, 2008).

Com a aplicação da IDCT obtém-se os coeficientes mel-cepstrais (Figura 3.12). No trabalho, determinou-se como 39 o número de coeficientes mel-cepstrais a serem extraídos de

cada quadro. Como exemplo, tem-se na Figura 3.13 os MFCC do quadro 8 do sinal de voz.

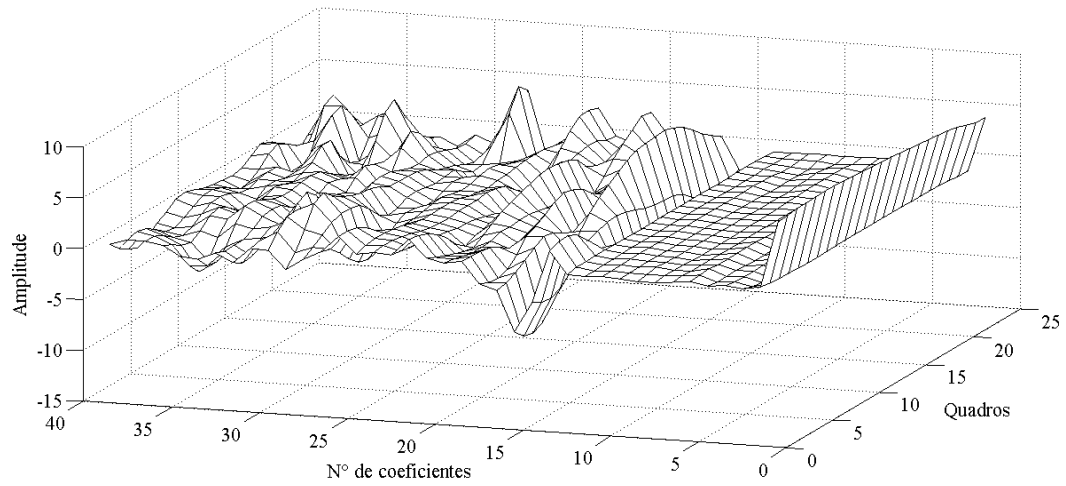

Figura 3.12: Coeficientes mel cepstrais do sinal de voz.
Fonte: Autoria própria.

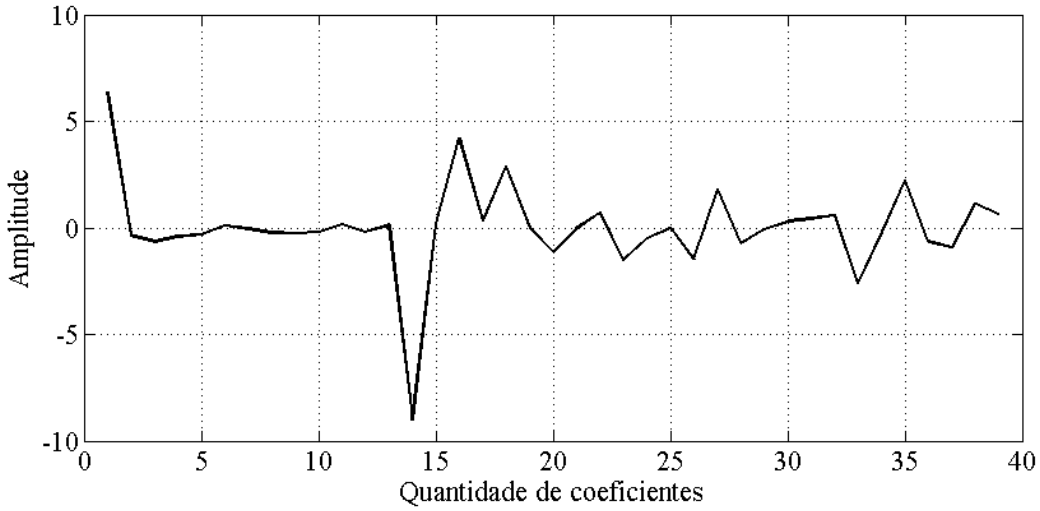

Figura 3.13: Coeficientes mel cepstrais do sinal de voz - Quadro 8.
Fonte: Autoria própria

3.2 HMM Aplicado ao Reconhecimento de Fala

Como já visto, o HMM é uma forma comum de representação da fala, modelando seus eventos por meio das distribuições de probabilidade de saída e a duração desses eventos por meio das probabilidades de transição. Assim, ele se torna importante para a modelagem do sinal já que elocuções de uma mesma palavra possuem diferentes durações (SILVA, 2009).

Em aplicações práticas, o treinamento de um HMM depende de várias sequências de observações independentes, devendo ser treinado com várias elocuções a fim de ser mais robusto.

A partir dos padrões extraídos na fase de processamento do sinal, tem-se o encadeamento desses no tempo e sua modelagem em uma máquina de estados finita, caracterizando assim o HMM. A partir desses padrões estima-se então o HMM de modo que cada elocução passa a ser representada por um sequência de estados.

No trabalho utilizou-se o *toolbox "Hidden Markov Model (HMM) Toolbox for Matlab"* de Murphy (2014), o qual possui as seguintes etapas:

1. Inicialização: representada pela função $[pi0, A0, B0, mu0, Sigma0]$ = *init_mhmm(data, N, M, cov_type)*, a qual tem por entrada os dados de treinamento (*data*), representados por uma matriz com os coeficientes mel-cepstrais de cada quadro; o número de estados do modelo (*N*); a quantidade de misturas de gaussianas (*M*) e do tipo de matriz, a qual pode ser diagonal, completa ou esférica. Da saída da função se obtem a distribuição de probabilidade incial de cada estado (*pi*); a matriz de distribuição de probabilidade de transição (*A*); a matriz de distribuição de probabilidade de observação (*B*); a matriz com os valores médios associados a cada mistura de gaussiana e a matriz as variâncias associadas às misturas de gaussianas.

2. Treinamento: representado pela função $[ll\_trace, pi, A, mu, sigma, B]$ = *mhmm_em(data, pi0, A0, mu0, sigma0, B0)*, a qual tem por entrada os dados do treinanmento (*data*) e as matrizes calculadas pela função de inicialização (*pi0, A0, B0, mu0, Sigma0*), e por saída as matrizes finais do HMM.

3. Reconhecimento: dado pela função $[Verossimilhança, Erros]$ = *mhmm_logprob(data_rec, pi, A, mu, sigma, B)*, na qual as entradas são os dados do sinal de voz a ser reconhecido (*data_rec*) e as matrizes do HMM, e as saídas são a verossimilhança, que estima a probabilidade da observação ter sido gerada pelo modelo e os erros, que indicam quando a verossimilhança não se adequa ao HMM.

Com o HMM definido, apresenta-se no próximo capítulo os resultados e discussões do trabalho.

4 Resultados e Discussões

Neste capítulo, serão apresentados os principais resultados obtidos com o sistema de reconhecimento de fala. Expõe-se, primeiramente, os dados obtidos para treinamento do HMM e, em sequência, os dados de validação, com as porcentagens de taxa de acerto de reconhecimento.

4.1 Base de dados

Devido as diferenças existentes entre as pronúncias de uma mesma palavra, que são maiores quando produzidas por locutores distintos, tem-se a necessidade de um número significativo de elocuções, de modo que o HMM seja capaz de absorver as variações existentes em seus padrões.

Para a criação da base de dados de um sistema de reconhecimento de fala independente de locutor, deve-se utilizar amostras de diferentes locutores, a fim de representar as mais variadas características dos possíveis usuários da aplicação desejada, tais como, sexo, idade, sotaque, entre outros.

Nste trabalho, a base de dados foi gerada por sete locutores, sendo quatro do sexo masculino e três do sexo feminino. Cada locutor gravou 50 amostras da palavra, totalizando 350 elocuções.

4.2 Treinamento do HMM

De acordo com Silva (2009), a etapa de treinamento é essencial para o sistema de reconhecimento de voz independente de locutor, influenciando significativamente no desempenho do sistema. Nesta fase, estima-se o HMM referente a palavra do vocabulário utilizado. Além disso, não existe uma regra padrão para determinar o número de estados necessários para a elocução, assim como para o número de misturas por estado. Estas decisões envolvem intuição

e familiaridade com os HMM, sendo necessário realizar testes experimentais com diferentes valores a fim de encontrar um número ótimo.

A palavra utilizada para o reconhecimento foi pato, que caracteriza-se por apresentar quatro fonemas (P-A-T-O), sendo dois consonantais (P-T) e dois vocálicos (A-O), duas sílabas e um espaço de silêncio entre a pronúncia.

Para treinamento do HMM foram coletadas 50 amostras de um locutor feminino que, após processadas conforme as etapas do Capítulo 3, puderam ser utilizadas para o treinamento e validação do modelo.

Inicialmente, os sinais utilizados como base foram divididos em dois grupos, um para treinamento, com 35 amostras, e um para validação do reconhecimento, com as 15 amostras restantes. Além disso, fez-se a divisão da elocução em duas partes, obtendo-se um HMM para reconhecer a sílaba PA e outro para reconhecer a sílaba TO. No desenvolvimento do código de reconhecimento, se definiu que o segundo HMM só seria verificado caso o primeiro reconhecesse a sílada determinada. Feito isso, realizou-se então, experimentos para encontrar o melhor número de estados e de misturas de gaussianas por estado.

A análise do HMM gerado pelo treinamento é realizada com a análise do índice de verossimilhança, a qual estima a probabilidade da entrada ter sido gerada pelo modelo (RAMOS, 2011). De acordo com Amaral (2014), maximizar a verossimilhança significa obter o modelo que tem a maior probabilidade ter gerado a amostra.

No primeiro treinamento realizado, fixou-se o número de estados igual a três, para que houvesse menos tempo de processamento, e se variou o número de misturas gaussianas, obtendo-se os valores de reconhecimento mostrados na Tabela 4.1.

Tabela 4.1: Teste 1 (3 estados)

Número de misturas de gaussianas	**Índice de verossimilhança**
3	-8.1568e+003
4	-4.1006e+003
5	-4.6917e+003
6	-1.3496e+003
7	73.5655
8	3.6195e+00
9	1.2599e+003
15	2.9137e+003

Como se nota, a partir de sete misturas de gaussianas o valor começa a ficar positivo, como se deseja para o sistema. Assim, escolheu-se primeiramente oito misturas de gaussia-

nas, já que testes com nove e quinze misturas tornaram o sistema mais lento. Mantendo-se esse número constante e variando-se o número de estados, obteve-se os valores da Tabela 4.2, referentes ao segundo treinamento.

Tabela 4.2: Teste 2 (8 misturas gaussianas)

Número de estados	**Índice de verossimilhança**
3	-1.4441e+003
4	1.8868e+003
5	3.2254e+003
6	3.4574e+003
7	7.4805e+003
8	5.6118e+003
9	5.0293e+003
10	4.2586e+003

Pela análise da Tabela 4.2, tem-se resultados que maximizam a verossimilhança a partir de quatro estados. Como o sistema começa a apresentar tempo significativo de processamento a partir de oito estados, adotou-se número de estados igual a sete.

Para a segunda metade da palavra, fazendo-se os mesmos testes, obteve-se as Tabelas 4.3 e 4.4.

Tabela 4.3: Teste 3 (3 estados)

Número de misturas de gaussianas	**Índice de verossimilhança**
3	-9.9222e+003
4	-8.8630e+003
5	-8.2679e+003
6	-5.1538e+003
7	-6.8443e+003
8	-4.7394e+003
9	-3.2153e+003
15	2.8886e+003

Para o segundo caso, adotou-se, também, número de misturas de gaussianas igual a quinze e número de estados igual a seis, já que a partir de sete, o tempo para processamento do sistema começou a aumentar significativamente.

Buscando-se confirmar o número de estados e de misturas de gaussianas, fez-se o teste de reconhecimento das 15 amostras destinadas a validação (mesmo locutor), obtendo-se 100% de acerto. A partir desse resultado, obteve-se os seguintes parâmetros do HMM:

Tabela 4.4: Teste 4 (15 misturas gaussianas)

Número de estados	**Índice de verossimilhança**
3	1.6277e+003
4	4.6844e+003
5	3.8573e+003
6	4.0170e+003
7	2.4220e+003
8	6.2933e+003
9	955.8597
10	9.4302e+003

- Matriz de probabilidade de transição entre estados:

  a) Para a primeira parte da elocução:

$$\begin{bmatrix} 0.9154 & 0.0846 & 0 & 0 & 0 & 0 \\ 0 & 0.7209 & 0.2791 & 0 & 0 & 0 \\ 0 & 0 & 0.7338 & 0.2662 & 0 & 0 \\ 0 & 0 & 0 & 0.6266 & 0.3734 & 0 \\ 0 & 0 & 0 & 0 & 0.8511 & 0.1489 \\ 0 & 0 & 0 & 0 & 0 & 1.0000 \end{bmatrix}$$

  b) Para a segunda parte da elocução:

$$\begin{bmatrix} 0.8592 & 0.1408 & 0 & 0 & 0 & 0 \\ 0 & 0.1379 & 0.8621 & 0 & 0 & 0 \\ 0 & 0 & 0.7976 & 0.2024 & 0 & 0 \\ 0 & 0 & 0 & 0.8924 & 0.1076 & 0 \\ 0 & 0 & 0 & 0 & 0.8889 & 0.1111 \\ 0 & 0 & 0 & 0 & 0 & 1.0000 \end{bmatrix}$$

- Distribuição de probabilidade de emissão de símbolos:

  a) Para a primeira parte da elocução:

$$\begin{bmatrix} 0.1293 & 0.1864 & 0.0975 & 0.1566 & 0.1647 & 0.1977 & 0.0678 \\ 0.2273 & 0.1970 & 0.0606 & 0.1212 & 0.2121 & 0.0758 & 0.1061 \\ 0.2126 & 0.1496 & 0 & 0.1181 & 0.0866 & 0.0236 & 0.4094 \\ 0.1882 & 0 & 0.4589 & 0.2117 & 0.0941 & 0.0470 & 0 \\ 0.3612 & 0.1100 & 0.2094 & 0.1257 & 0.0209 & 0.1728 & 0 \\ 0 & 0 & 0 & 0.3707 & 0 & 0.0490 & 0.5804 \end{bmatrix}$$

b) Para a segunda parte da elocução:

$$\begin{bmatrix} 0.0451 & 0.1352 & 0.5211 & 0.0958 & 0.1859 & 0.0056 & 0.0113 \\ 0.0172 & 0.0345 & 0.2241 & 0 & 0 & 0.7241 & 0 \\ 0.0931 & 0.1255 & 0.0769 & 0.2348 & 0.1296 & 0.2105 & 0.1296 \\ 0 & 0.1124 & 0.0201 & 0.2209 & 0.1853 & 0.2364 & 0.2249 \\ 0.1397 & 0.2721 & 0.2132 & 0.1397 & 0 & 0.0515 & 0.1838 \\ 0.0286 & 0.3810 & 0.3333 & 0.0095 & 0.0571 & 0 & 0.1905 \end{bmatrix}$$

- Distribuição de estado inicial:

a) Para a primeira e segunda parte da elocução, sendo $\{\ldots\}^T$ o operador transposto:

$[1\ 0\ 0\ 0\ 0\ 0]^T$

4.3 Validação Cruzada do HMM

Com o objetivo de confirmar os valores das matrizes que representam o HMM, fez-se a validação cruzada proposta por HAYKIN e VEEN (2001), a qual é uma forma de validar o modelo com um conjunto de dados diferentes do usado para estimar os parâmetros. Para isso, o conjunto de dados disponível foi dividido em um conjunto de treinamento e um conjunto de teste.

Para realização da validação cruzada foram utilizadas as 50 amostras do locutor utilizado na seção 3.2, sendo 35 para treinamento e 15 para validação, obtendo-se os resultados conforma a Tabela 4.5.

Pela análise dos resultados, tem-se que o modelo criado pelas amostras de treinamento de 1 à 35 e 13 à 47 apresentaram as maiores médias de índice de verossimilhança.

Tabela 4.5: Testes de validação cruzada - 35 amostras para treinamento e 15 para validação

Amostras para Treinamento	**Índice de verossimilhança 1ª metade**	**Índice de verossimilhança 2ª metade**
1-35	9.1099e+003	2.5896e+003
2-36	6.6879e+003	3.8349e+003
3-37	5.1210e+003	2.2149e+003
4-38	9.3404e+003	2.2964e+003
5-39	7.4492e+003	2.5963e+003
6-40	1.1045e+004	4.5963e+003
7-41	5.4194e+003	4.9035e+003
8-42	5.1908e+003	2.6611e+003
9-43	5.1618e+003	2.7548e+003
10-44	3.3460e+003	6.3841e+003
11-45	3.6642e+003	5.1443e+003
12-46	2.7273e+003	3.9248e+003
13-47	4.3916e+003	7.1912e+003
14-48	1.2396e+003	5.3197e+003
15-49	1.3534e+003	6.5743e+003
16-50	2.9332e+003	6.1618e+003

Ao se comparar as matrizes desses dois modelos, pode-se confirmar os valores das matrizes do modelo da seção 4.2, a qual, a partir de então, foi utilizada para todos os testes.

4.4 Validação com Outros Locutores

Para validação do sistema com diferentes locutores, utilizaram-se as amostras dos outros seis locutores (quatro masculinos e dois femininos). Após serem processadas, fez-se o teste de reconhecimento, obtendo-se os resultados mostrados na Tabela 4.6.

Tabela 4.6: Teste de reconhecimento para locutores diferentes

Locutor	**Taxa de acerto**
Masculino 1	96 %
Masculino 2	74 %
Masculino 3	100 %
Masculino 4	94 %
Feminino 1	92 %
Feminino 2	96 %

Como se nota, em cinco casos os resultados foram superiores a 90%, mesmo com a utilização de apenas um locutor para treinamento do HMM. Isso demonstra que o sistema reage a locutores diferentes, mesmo que não tenha sido treinado com sua voz especificamente, o que

atinge o objetivo do sistema de reconhecer determinada elocução independetemente do locutor.

Porém, visando aumentar a taxa de acerto do sistema, se fez a divisão das 50 elocuções de cada locutor em dois grupos, a fim de treinar o sistema com sinais de voz de todos os locutores. Para isso, foram selecionadas 15 amostras de cada locutor para treinamento e 35 amostras para validação, totalizando 105 amostras para treinamento e 245 para validação. Nesse teste o sistema apresentou taxa de acerto de 98%, o que confirma os melhores resultados do HMM quando treinado com elocuções que apresentam caracterísiticas diferentes.

Diante dos resultados obtidos, tem-se no próximo capítulo as considerações finais do trabalho.

5 Conclusão

Neste trabalho, teve-se por objetivo o desenvolvimento de um sistema de reconhecimento de fala capaz de reconhecer uma palavra isolada com a utilização dos HMM.

Inicialmente foram estudados aspectos importantes na área de processamento de sinais e reconhecimento de padrões, fazendo-se um levantamento bibliográfico dos métodos utilizados na área do reconhecimento de fala e definindo-se as etapas a serem seguidas até o desenvolvimento final do sistema de reconhecimento.

No decorrer do levantamento bilbiográfico, se decidiu pela utilização dos MFCC como a técnica para extração dos padrões dos sinais de voz, utilizados no treinamento e validação do HMM.

Em seguida, se teve o desenvolvimento no software MATLAB do sistema de reconhecimento de fala independente de locutor utilizando HMM contínuos.

Para cumprir todas as etapas da obtenção dos MFCC foi necessário utilizar um *toolbox*, o qual demonstrou atender às necessidades iniciais do projeto. Também foi utilizado um *toolbox* para os HMM, formado pelas etapas de inicialização, treinamento e reconhecimento. Em trabalho futuros, espera-se conseguir independência desses *toolboxes*, já que muitas vezes eles apresentam conflitos na execução do sistema, não convergindo para o reconhecimento de forma adequada.

A confecção da base de dados demonstrou-se como uma etapa importante para aumentar a taxa de acerto do sistema. Foram utilizados 7 locutores (4 masculinos e 3 femininos) os quais apresentaram variação de pronúncia e de ritmo ao pronunciar a elocução. Nos primeiros testes, o sistema foi treinado e validado com amostras de um mesmo locutor, gerando-se o HMM a ser implementado. Com os parâmetros definidos, testou-se esse HMM com elocuções dos locutores restantes, conseguindo-se uma taxa de acerto inferior a 90% em apenas um caso, o que pode ser considerado satisfatório. Para confirmar o modelo gerado fez-se a utilização da validação cruzada que resultou em parâmetros do HMM semelhantes ao já obtido.

Na sequência, visando tornar o sistema mais robusto, fez-se o treinamento do HMM

com elocuções dos 7 locutores. Os 50 sinais de fala de cada locutor foram divididos em 15 amostras para treinamento e 35 para validação. O objetivo dessa etapa foi fazer com que o sistema fosse alimentado com as mais diversas formas de pronúncia da palavra, elevando sua taxa de acerto para 100%, o que foi atingido. Isso demonstra que o HMM deve ser treinado com a maior quantidade de elocuções que representem as diversas formas de sua pronúncia, reagindo de forma mais satisfatória conforme a quantidade de dados que recebe.

Além disso, se fez o teste do sistema com elocuções semelhantes à de interesse, tais como “pata” e “gato”, obtendo-se resultados de 12% de erro, o que demonstra um bom funcionamento do sistema, mas com a necessidade de ajustes.

Na continuação do trabalho, se espera validar o modelo gerado no último teste. Para isso, será necessário a aquisiçao de sinais de fala de outros locutores, a fim de verificar se o sistema fará o reconhecimento correto.

Assim, se entende que o projeto desenvolvido é o início da tecnologia do reconhecimento de fala, podendo ser aprimorado para fazer o reconhecimento do locutor e/ou acionamento de sistemas de interesse, como proposto no início.

Referências Bibliográficas

AMARAL, G. J. A. Teoria da regressão: Texto introdutório. Disponível em: `www.de.ufpe.br/~gjaa/Introreg.doc`. Acesso em 05 de out. 2014.

ATAL, B. S.; HANAUER, S. L. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am*, v. 50, n. 2, p. 637–655, 1971.

BAKER, J. K.; BAHL., L. R. Some experiments in automatic recognition of continuous speech. *Proceedings of the 11th Annual IEEE Computer Society Conference*, p. 326–329, 1975.

BAUM, L. E.; EAGON, J. A. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, n. 3, p. 360–363, 1967.

BAUM, L. E.; PETRIE, T. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, n. 6, p. 1554–1563, 1966.

BAUM, L. E. et al. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, n. 1, p. 164–171, 1970.

BOGERT, B.; HEALY, M.; TUKEY, J. The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Proceedings of the Symposium on Time Series Analysis*, p. 209–243, 1963.

BOUROUBA E-H., B. M. D. R. Isolated words recognition system based on hybrid approach dtw/ghmm. *Informatica, An International Journal of Computing and Informatics*, v. 30, n. 3, p. 373–384, 2006.

BRAGA, P. de L. *Reconhecimento de voz dependente de locutor utilizando Redes Neurais Artificiais*. 85 p. Monografia (Engenharia da Computação) — Universidade de Pernambuco, 2006.

BROOKES, M. Voicebox: Speech processing toolbox for matlab. Disponível em: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`. Acesso em 15 de set. 2014.

CARVALHO, G. P. S. de; SANTOS, T. M. dos. Biometria: Impressão vocal. Disponível em: `http://www.gta.ufrj.br/grad/09_1/versao-final/impvocal/index.html`. Acesso em 05 de out. 2014.

DAVIS, K. H.; BIDDULPH, R.; BALASHEK, S. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 1952.

DELLER Jr., J. R.; HANSEN, J. H. L.; PROAKIS, J. G. *Discrete time processing of speech signals*. New York: MacMillan Publishing Company, 1993.

DUDLEY, H. The vocoder. *Bell Labs Record*, v. 17, p. 122–126, 1939.

DUDLEY, R. R. R. H.; WATKINS, S. A. A synthetic speaker. *J. Franklin Institute*, v. 227, p. 739–764, 1939.

ESPINDOLA, L. da S. *Um estudo sobre Modelos Ocultos de Markov*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio Grande do Sul - Faculdade de Informática, 2009.

FLETCHER, H. The nature of speech and its interpretations. *Bell Syst. Tech. J.*, v. 1, p. 129–144, 1922.

HAYKIN, S.; VEEN, B. V. *Sinais e Sistemas*. Porto Alegre: Bookman, 2001.

ITAKURA, F.; SAITO, S. A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Communications in Japan*, v. 53A, p. 36–43, 1970.

JELINEK, F. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, n. 4, p. 532–556, 1976.

JUANG, B.; RABINER, L. R. Automatic speech recognition: A brief history of the technology development. Georgia Institute of Technology, Atlanta, 2004.

MICROSOFT, M. D. Say hello to cortana, your truly personal assistant. Disponível em: `http://www.microsoft.com/en-us/mobile/campaign-cortana/`. Acesso em 08 de nov. 2014.

MURPHY, K. Hidden markov model (hmm) toolbox. Disponível em: `http://www.cs.ubc.ca/~murphyk/Bayes/PreMIT/hmm.html`. Acesso em 08 de set. 2014.

OLIVEIRA, M. P. B. *Verificação automática do locutor, dependente do texto, utilizando sistemas híbridos MLP/HMM*. Dissertação (Mestrado) — Instituto Militar de Engenharia, 2001.

OPPENHEIM, A. V.; SCHAFER, R. W. *Processamento de tempo discreto de sinais*. 3. ed. São Paulo: Perason Education do Brasil, 2012.

PICONE, J. Signal modeling techniques in speech recognition. *Proceedings of IEEE*, v. 81, n. 8, p. 1215–1247, 1993.

RABINER, L.; JUANG, B.-H. *Fundamentals of Speech Recognition*. New Jersey: Englewood Cliffs: Prentice Hall, 1978.

RABINER, L.; JUANG, B. H. *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall, 1993.

RABINER, L.; SCHAFER, R. W. *Digital processing of speech signals*. New Jersey: Prentice Hall, 1978.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, n. 77, p. 257–286, 2002.

RABINER, L. R.; SCHAFER, R. W. Introduction to digital speech recognition. *Foundations and Trends in Signal Processing*, v. 1, n. 1-2, p. 1–194, 2007.

RAMOS, M. Cadenas ocultas de markov aplicadas al reconocimiento de voz. Universidade de la República Uruguay, 2011.

SILVA, A. G. da. *Reconhecimento de voz para palavras isoladas*. 52 p. Monografia (Engenharia da Computação) — Universidade Federal de Pernambuco, Recife, 2009.

SOTERO Jr., R. S. *Investigação de um ambiente call center utilizando reconhecimento de fala*. 42 p. Monografia (Ciência da Computação) — Universidade Federal de Pernambuco, Recife, 2011.

TAFNER, M. A. *Reconhecimento de palavras faladas isoladas usando redes neurais artificiais*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 1996.

TEVAH, R. T. *Implementação de um sistema de reconhecimento de fala contínua com amplo vocabulário para o português brasileiro*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 1996.

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

YNOGUTI, C. A. *Reconhecimento de fala contínua usando Modelos Ocultos de Markov*. Tese (Doutorado) — Universidade Estadual de Campinas, 1999.

YOMA, N. B. *Reconhecimento automático de palavras isoladas: estudo e aplicação dos métodos determinístico e estocástico*. Dissertação (Mestrado) — Universidade Estadual de Campinas, 1993.

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

APÊNDICE A – Códigos

Neste anexo, apresentam-se as etapas do processo de reconhecimento implementadas no *software* Matlab. Para criar o banco de filtros na escala Mel, utilizou-se o *Voicebox* de Brookes (2014) e para o desenvolvimento do HMM, o *toolbox* escrito por Murphy (2014).

O código apresentado se refere ao reconhecimento da primeira sílaba da elocução, sendo semelhante para a segunda parte.

```
%*******************************************************%
% SISTEMA DE RECONHECIMENTO DE FALA USANDO HMM E MFCC %
% Parte 1 - 1ª parte da elocução
%*******************************************************%

clear all; clc; close all
Fs = 22050;

for i=1:105
%*******************************************************%

% 1 AQUISIÇÃO DO SINAL
    s = sprintf('pato%d.wav',i);
    sinal_base = wavread(s);

%*******************************************************%

% 2 PRÉ-PROCESSAMENTO

% 2.1 Retirada do nível médio
sinalAC = sinal_base- mean(sinal_base);

% 2.2 Normalização da amplitude
normalizado = sinalAC/max(abs(sinalAC));

% 2.3 Corte do Sinal
```

```
sinal_cortado1 = [normalizado(1:5512)];

%*****************************************************%

% 3 EXTRAÇÃO DE PADRÕES

% 3.1 Filtragem de pré-ênfase
B = [1 -0.95]; % coeficientes do filtro H(z) = 1 - azˆ-1
% alpha varia entre 0.95 e 0.98
pre_enfase1= filter(B,1,sinal_cortado1);

% 3.2 Divisão do sinal em frames e janelamento - janela de Hamming
frame_len = 256; % duraçao do frame de aprox. 11,6 ms
n = length(pre_enfase1);
num_frames = floor(n/frame_len); % arredonda o num_frames p/ o inteiro
% mais próximo de menos infinito. = 21 quadros
win_length = 256; % N - número de amostras por frame
win_overlap = 80; % M - número de amostras sobrepostas
janela = hamming(win_length);

for k = 1 : num_frames

    % Aplicando Hamming em cada quadro
    % 1°: Dividir o sinal de 256 em 256 amostras
    frame = pre_enfase1((((k-1)*frame_len+1)-(k-1)*win_overlap)...
    :((frame_len*k)-((k-1)*win_overlap)));
    matriz_frames(:,k) = frame; % matriz coluna dos frames
    % 3°: Aplicando Hamming em cada coluna da matriz
    hamming = matriz_frames(:,k).*janela; % hamming em cada frame
    % 4°: Invertendo a matriz - 20X220
    matriz_hamming(k,:) = hamming; % matriz linha do janelamento

    % Aplicando FFT em cada quadro
    % Comprimento de cada linha da matriz onde se aplicará a FFT
    L_hamming = length(matriz_hamming(k,:));
    % Potência de 2 para representar o comprimento da FFT
    nfft = nextpow2(L_hamming);
    % Aplicando FFT em cada linha da matriz
    hamming_fft = fft(matriz_hamming(k,:),2ˆnfft);
    % Salavando todas as FFT em uma matriz, mantendo as partes únicas
    matriz_hamming_fft(k,:) = hamming_fft(1:(floor((2ˆnfft+1)/2)));
    matriz_hamming_fft(k,:) = matriz_hamming_fft(k,:)/(L_hamming/2);
```

```
end

% 3.3 Banco de filtros Mel

p = 39; % número coeficientes = número de filtros
n = 257; % comprimento da FFT
fl = 0; % extremidade inferior do menor filtro como fração de fs
fh = 0.5; % extremidade superior do maior filtro como fração de fs
% fl = 0 e fh = 0.5 são valores padrão
w = 't'; % forma do filtro: triangular

[x,mc,na,nb] = melbankm(p,n,Fs);
% x é matriz esparsa, para transformar em matriz completa:
% x_full = full(x);

%*********************************************************%

% 4 ANÁLISE CEPSTRAL

filtro = x';

% 4.1 Log
for k = 1 : num_frames
    f = matriz_hamming_fft(k,:);
    pot = log((abs(f).^2)*filtro);
    cepstro1(:,k) = pot;
end

% 4.2 IDCT

MFCC1= idct(cepstro1);

matriz_MFCC1(:,:,i) = [MFCC1];

end

%*********************************************************%
%                 MODELO OCULTO DE MARKOV
%*********************************************************%

dados =  MFCC1; % dados para treinamento

N = 7; % número de estados
```

```
M = 8; % número de misturas de gaussianas
cov_type = 'diag';
% 1 - modelo left-right

% 1. Inicialização
[pi0, A0, B0, mu0, sigma0] = init_mhmm(dados, N, M, cov_type, 1);

% 2. Treinamento
[ll_trace, pi1, A1, mu1, sigma1, B1] = mhmm_em(dados, pi0, A0, mu0,...
sigma0, B0);

% Reconhecimento
[Verossimilhanca1,Erro1] = mhmm_logprob(dados_reconhecimento, pi1, ...
A1,mu1, sigma1, B1);
```