1
Introdução

Esta tese visa apresentar novas propostas de combinação de classificadores, aplicados em sub-bandas, buscando melhorar a taxa de acerto do reconhecimento de locutor quando realizado na presença de condições adversas, tais como ruídos no sinal de fala.

Neste capítulo, é inicialmente apresentada na Seção 1.1 uma breve revisão de sistemas de reconhecimento de locutor e uma descrição da forma como o ruído ambiente afeta o seu desempenho. Em seguida, na Seção 1.2, são apresentados o objetivo da tese, a motivação e algumas considerações sobre as novas abordagens relacionadas às propostas deste trabalho. Finalmente, a Seção 1.3 descreve a organização dos capítulos da tese.

1.1
Sistemas de Reconhecimento de Locutor

A área de processamento de voz denominada reconhecimento de locutor visa reconhecer locutores a partir das propriedades da sua voz. Essa área pode ser subdividida em duas outras: identificação de locutor [1], que tem por objetivo identificar um determinado locutor dentre um grupo de locutores, e verificação de locutor [2], que consiste em verificar a associação da voz a uma determinada pessoa. O reconhecimento pode ser dependente do texto, ou seja, o reconhecimento utiliza obrigatoriamente as mesmas palavras usadas no projeto do reconhecedor; ou independente do texto, quando o reconhecimento independe de quais palavras foram usadas no projeto do sistema. Geralmente o desempenho (em termos da taxa de acerto) de sistemas de reconhecimento dependentes do texto tende a ser melhor do que o de sistemas de reconhecimento independentes do texto, já que no primeiro caso os dados de projeto e os de teste dos sistemas são similares. Os locutores podem assumir um comportamento cooperativo [3] ajudando o sistema a realizar o reconhecimento, por meio de falas pausadas, bem pronunciadas; ou não cooperativo, quando, por exemplo, os locutores não sabem que estão sendo reconhecidos e por esse motivo não há a preocupação de gerar falas organizadas e bem articuladas. Existem várias aplicações de grande importância para as técnicas de reconhecimento de locutor, como por exemplo, o

auxílio às investigações criminais envolvendo vozes gravadas, a realização de transações bancárias, o controle de acesso (sensor biométrico) a dispositivos e redes de dados, determinadas aplicações militares e auxílio a deficientes.

O projeto do sistema de reconhecimento de locutor, como mostrado na Figura 1, envolve inicialmente a extração de atributos (parâmetros, características) das amostras do sinal de voz. Os atributos possuem informação da identidade do locutor e são propriedades intrínsecas do sinal, como por exemplo, o mel-cepestro, que é obtido de um banco de filtros [4] espaçados uniformemente em uma escala de frequências chamada Mel [5]. A etapa seguinte consiste na classificação dos atributos que é realizada por sistemas classificadores. Esses sistemas operam gerando um modelo que visa representar estatisticamente o locutor. Um classificador que tem sido bastante utilizado em aplicações de reconhecimento independente do texto, devido ao seu bom desempenho, é o GMM (*Gaussian Mixture Models*) [6]. Esse sistema cria uma representação estatística usando combinações lineares de funções gaussianas.

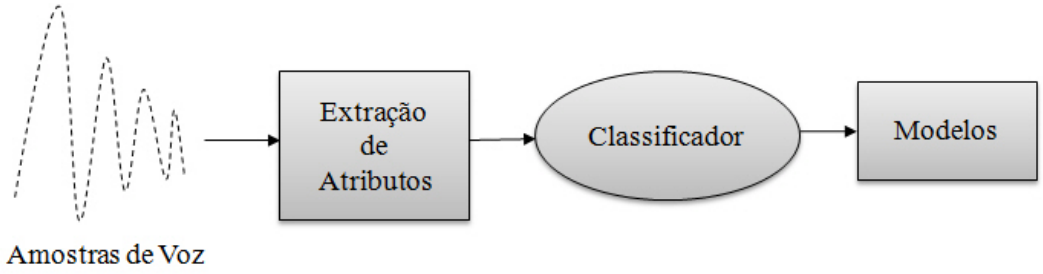


Figura 1 – Etapa de treinamento de um sistema automático de reconhecimento de locutor

O teste de um reconhecedor envolve a extração de atributos das amostras do sinal de voz do mesmo tipo daqueles que são usados em seu projeto. Em seguida, é feita uma decisão dentre qual modelo é mais apropriado a esse conjunto de atributos, como mostrado a seguir na Figura 2. O locutor cujo modelo tem maior probabilidade de gerar o conjunto de atributos testados é o reconhecido. Usualmente essa comparação é realizada através de uma medida estatística chamada função de verossimilhança, que mede probabilisticamente o quanto um conjunto de padrões de teste se assemelha ao conjunto de padrões que gerou o modelo armazenado [1].

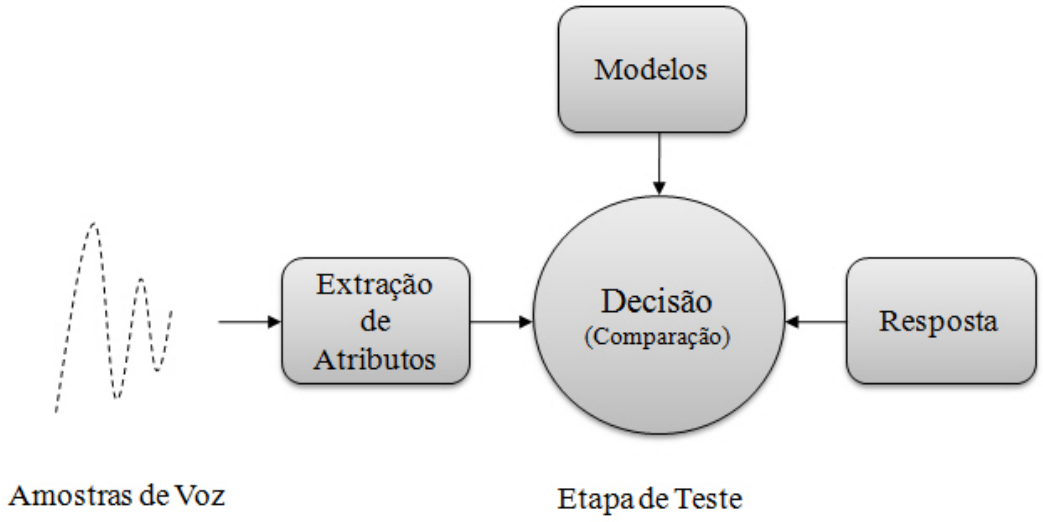


Figura 2 – Etapa de teste do sistema automático de reconhecimento de locutor

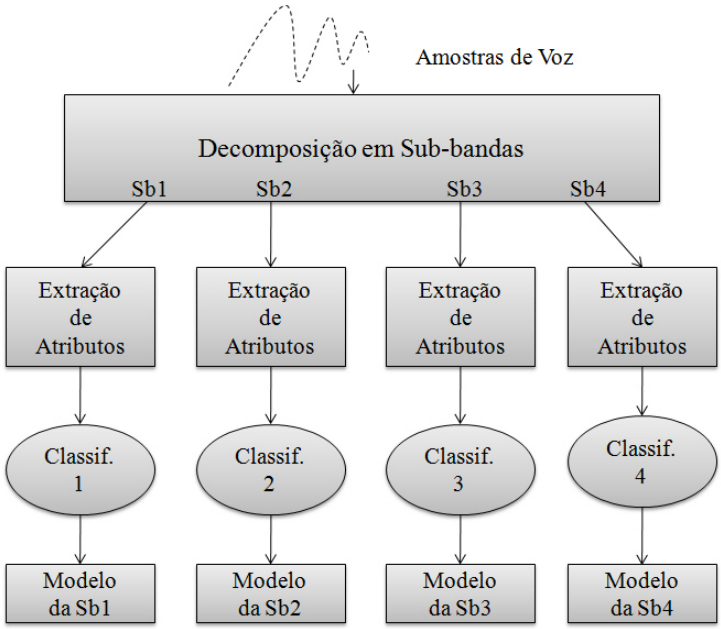
As abordagens empregadas em sistemas de reconhecimento de locutor, que tendem a apresentar os melhores resultados em termos de desempenho, são aquelas que procuram imitar o comportamento biológico [1]. Vários estudos são desenvolvidos com o principal objetivo de gerar modelos matemáticos que melhor representem esse comportamento. A partir desses estudos, foi criada, por exemplo, a escala Mel usada na obtenção dos atributos MFCC (*Mel Frequency Cepstral Coefficients*) [4]. Outro exemplo são os filtros cocleares utilizados na obtenção dos atributos ZCPA (*Zero Crossings with Peak Amplitudes*) [7],[8].

Os sistemas de reconhecimento têm o seu desempenho afetado por uma variedade de fatores, tais como o ruído no sinal de fala, o número de locutores, distorções provocadas pelo canal de comunicação e descasamentos acústicos provocados pelas diferentes condições de treinamento (projeto) e de teste do reconhecedor [6], [9], [10]. Essas diferentes condições podem estar relacionadas, por exemplo, a treinar o sistema usando a voz proveniente da telefonia fixa e a testar com a voz originada da telefonia celular. Nota-se que o canal de comunicação da telefonia fixa e o canal de telefonia móvel tendem a distorcer distintamente o sinal de voz, podendo provocar descasamento no reconhecimento. Além disso, os sons de fábrica, de carros e até de várias pessoas falando ao mesmo tempo são fontes de ruídos que pioram o reconhecimento. O uso de diferentes tipos de microfones e as condições fisiológicas do aparelho vocal dos locutores também tendem a degradar o desempenho do reconhecimento. A robustez do reconhecedor a esses fatores indesejados, já que são os mais encontrados na prática, é de extrema importância.

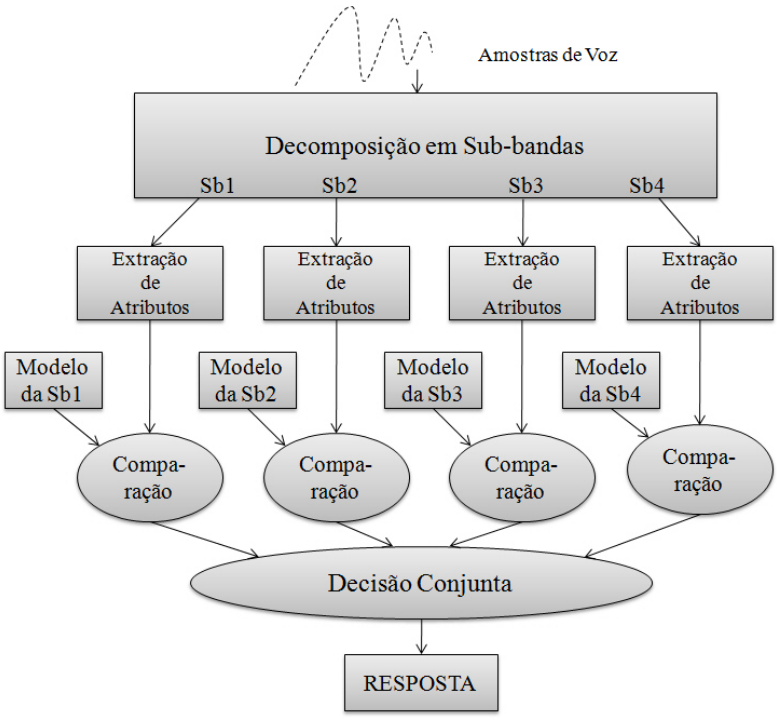
O uso de técnicas de pré-processamento tem ajudado a minimizar o efeito indesejado provocado pelos fatores que prejudicam o desempenho [9], [11]-[27]. Uma dessas técnicas, chamada de CMS (*Cesptral Mean Subtraction*), foi mostrada em [9] e consiste em subtrair de cada atributo uma média que é calculada usando uma coleção de atributos afetados pelo canal de comunicações. Ou seja, consiste em simplesmente remover de cada atributo a sua média temporal, que é considerada uma estimativa destes efeitos. Além disso, a utilização de novos tipos de atributos [28]-[37], que visem representar melhor o sinal de voz, como por exemplo, o parâmetro de Hurst (H) [29], têm melhorado o desempenho do sistema de reconhecimento em determinadas aplicações. O emprego dessas técnicas geralmente, requer que a voz e a degradação obedeçam a algumas restrições de acordo com a sua natureza, tais como a descorrelação, a aditividade, entre outras. Além disso, a utilização dessas abordagens tende a ser específica ao tipo de aplicação. Por esse motivo, não é prático o uso dessas estratégias em situações em que nada se sabe sobre o tipo e a natureza da degradação presente no sinal.

Abordagens recentes têm usado a decomposição do sinal de voz em sub-bandas (ou faixas) de frequências e aplicado um classificador aos dados referentes a cada um desses intervalos espectrais [38]-[40]. Nessas estratégias, o locutor é representado por um conjunto de modelos, sendo que cada um deles é modelado a partir dos dados extraídos de uma das sub-bandas. A Figura 3(a) ilustra um esquema para a decomposição em quatro sub-bandas. A etapa de teste usa a verossimilhança conjunta, que é calculada fazendo uma combinação das verossimilhanças individuais resultantes de cada um desses modelos, como mostrado na Figura 3(b). Essas estratégias exploram as vantagens da construção de modelos a partir da decomposição do sinal em sub-bandas de frequências, já que as contribuições indesejadas, como os ruídos, podem afetar mais certas regiões do espectro do sinal de voz do que outras. Recentemente, as pesquisas de técnicas baseadas em decomposição em sub-bandas vêm tendo bastante ênfase não só em reconhecimento de locutor, mas também em outras áreas como em processamento de imagens [41], [42]. Uma técnica de reconhecimento proposta em [38] usa, como critério de combinação, a soma das respostas de cada classificador a fim de obter a decisão conjunta final. Isso equivale a aplicar pesos lineares e unitários à saída dos classificadores. A proposta apresentada em [39] é semelhante à anterior, mas utiliza somente as quatro sub-bandas de maior energia.

As abordagens baseadas em sub-bandas têm a vantagem de permitir o desenvolvimento de técnicas que não sejam específicas para um determinado tipo de degradação presente no sinal.



(a)



(b)

Figura 3 – Sistema de reconhecimento usando múltiplos classificadores em sub-bandas. (a) Construção dos modelos. (b) Teste do sistema

Além do reconhecimento de locutor, que utiliza a voz, existem outras aplicações que requerem o emprego de múltiplos classificadores. Um exemplo, apresentado em [43], baseia-se em reconhecer pessoas por meio da combinação

das informações provenientes da voz com as da face. Quanto às técnicas de combinação de classificadores, algumas estratégias [44]-[56] baseiam-se em critérios probabilísticos em que cada classificador recebe uma "nota" de confiança. Essa nota expressa o quão provável a resposta do classificador representa a verdade. Ou seja, se um classificador tem uma nota de confiança de 80% (em uma escala de 0% a 100%), a sua resposta contribui mais para a decisão conjunta final envolvendo todos os classificadores. Um problema dessas técnicas é a necessidade de execução, durante o projeto do sistema de reconhecimento, de vários testes iniciais dos classificadores, exigindo um número adequado de dados de entrada ou, na falta parcial destes, o emprego de reamostragens desses dados, de forma a se obter as estatísticas necessárias para a operação adequada da estratégia de combinação. Ressalta-se também que essas técnicas são difíceis de serem modeladas. O emprego de estratégias de combinação mais simples, como as baseadas em técnicas lineares apresentadas em [38] e [39], e as que estão sendo propostas nesta tese, são fundamentais para aplicações práticas de reconhecimento de locutor. É importante lembrar que os sistemas de reconhecimento, que utilizam múltiplos classificadores, requerem um certo esforço computacional para a realização da tarefa de combinação e para as etapas de extração de atributos e classificação. O esforço computacional total tende a ser elevado, o que pode comprometer a utilização desses sistemas em determinadas aplicações. Adicionalmente, estas técnicas podem não apresentar resultados satisfatórios quando operam em situações de desconhecimento acerca do tipo e natureza da degradação presente no sinal de teste, o qual poderá ainda possuir uma quantidade insuficiente de amostras. Desta forma, pretende-se que as novas propostas, sem necessitar de elevados níveis de processamento, sejam eficientes, mesmo com o desconhecimento do tipo de ruído presente no sinal.

1.2
Considerações Iniciais Sobre as Novas Propostas

O principal objetivo desta tese consiste em desenvolver novas estratégias de combinação de classificadores aplicados em sub-bandas, visando melhorar o desempenho de sistemas de reconhecimento de locutor na presença de condições adversas, tais como ruídos no sinal de fala. Este trabalho é concentrado na área de

identificação automática de locutor independente do texto, com sinal de fala degradado por ruído. O ruído é a maneira mais frequente de degradar um sistema de reconhecimento, uma vez que está usualmente presente no ambiente.

A pesquisa desenvolvida busca melhorar a taxa de reconhecimento em aplicações onde nada se sabe sobre o tipo e a natureza do ruído. As pesquisas mais recentes [26] têm se concentrado neste tipo de situação, já que é uma das mais comuns, haja vista a variedade de ruídos, e tem-se mostrado como uma das mais difíceis de obter melhorias na taxa de acerto. Note-se, portanto, a importância do estudo desenvolvido nessa tese. Nesse contexto, são propostas e analisadas novas estratégias de combinação que aplicam pesos não-uniformes às respostas dos classificadores empregados em cada sub-banda. Isso porque apesar da técnica Soma ter sido apontada como a melhor estratégia [38], dentre as simuladas para aplicações que envolvam a voz de teste contaminada por ruído passa-faixa, pretende-se mostrar nesta tese que melhorias na taxa de acerto podem ser alcançadas por meio da utilização de pesos não-uniformes. A motivação é favorecer as respostas associadas às sub-bandas mais relacionadas à informação da identidade do locutor, visando melhorar o desempenho da identificação. Por exemplo, em uma primeira proposta, os pesos são calculados diretamente das energias do sinal em cada sub-banda. Isso porque a informação da identidade do locutor está principalmente concentrada em algumas frequências, especialmente nas mais baixas [4],[39]. Consequentemente, as contribuições devido a essas frequências tendem a ser maiores do que as de outras frequências mais relacionadas a ruídos. Note-se que não estão sendo empregadas técnicas de pré-processamento para a extração de ruídos (*denoising*).

As novas propostas evitam o emprego de restrições quanto à natureza e o tipo de degradação presente no sinal, visando facilitar a utilização prática dos sistemas de reconhecimento robusto. Ao longo da pesquisa, atenção especial tem sido dada no sentido de procurar simplificar o esquema funcional dos sistemas de reconhecimento. As propostas apresentadas são simuladas com vozes corrompidas por diferentes tipos de ruídos comumente encontrados no dia-a-dia. Também é utilizada uma base de vozes que tem sido considerada na literatura internacional. Isso tem por objetivo avaliar eficientemente o desempenho das propostas.

O trabalho aqui apresentado representa uma contribuição original importante e útil às aplicações que necessitam de identificação de locutor robusta e independente do texto.

1.3
Organização da Tese

Esta tese está organizada em oito capítulos. A Introdução mostra a importância do uso das técnicas de reconhecimento de locutor em diversas aplicações, os principais problemas nessa área e a motivação para o desenvolvimento de novas propostas. Além disso, fornece uma breve revisão sobre as técnicas de reconhecimento, apresenta o principal objetivo desta pesquisa, fornece algumas considerações iniciais sobre as novas propostas e apresenta a organização da tese.

O Capítulo 2 descreve resumidamente exemplos de técnicas de extração de atributos e de classificadores empregados em reconhecimento de locutor em ambientes ruidosos. Essas informações são importantes para o entendimento dos capítulos seguintes.

O Capítulo 3 inicialmente descreve as técnicas de combinação de classificadores em sub-bandas presentes na literatura. Em seguida é proposta uma estratégia para a escolha dos pesos aplicados às respostas dos classificadores nas sub-bandas, visando aumentar a taxa de acerto da identificação na presença de ruído. Essa proposta calcula os pesos a partir das energias dos sinais nas sub-bandas.

O Capítulo 4 apresenta uma outra proposta de combinação das respostas dos classificadores em sub-bandas, visando aumentar ainda mais a taxa de acerto do reconhecimento realizado sob condições adversas. Uma vez que na fase de teste o tipo de ruído pode mudar dependendo da aplicação, essa técnica utiliza o cálculo do espaço nulo com a finalidade de obter conjuntos de pesos que dependam da informação das energias nas bandas. A estratégia permite que os pesos sejam alternados na fase de teste, permitindo que o reconhecimento seja realizado com o melhor conjunto das ponderações. Essa proposta representa uma melhoria da anterior, porém a um custo computacional mais elevado já que envolve o cálculo de uma base para o espaço nulo, o armazenamento dos conjuntos de pesos para a utilização na fase de teste, e a alternância das ponderações nessa fase. Adicionalmente, é analisada a aplicação de treinamento em múltiplas condições [26], na nova técnica de reconhecimento que utiliza o espaço nulo. Busca-se melhorar o desempenho, em termos de taxa de acerto, através de uma compensação no modelo para o efeito do ruído. Nessa estratégia,

os classificadores são treinados com ruído. Uma proposta adicional, visa investigar como o uso de atributos dinâmicos (delta e delta-delta) [6] pode contribuir para a melhoria do reconhecimento utilizando a proposta baseada no espaço nulo. Os atributos dinâmicos fornecem informações adicionais sobre a identidade do locutor, provenientes das propriedades dinâmicas presentes na voz.

No Capítulo 5, é proposta a utilização de um tipo de atributo usado em reconhecimento de voz robusto, cuja técnica de extração utiliza os coeficientes de autocorrelação do sinal de voz [31].

Finalizando, no Capítulo 6, são apresentadas as principais conclusões da tese e as sugestões para a continuação deste trabalho de pesquisa.